Integrative Analysis of Genetic Associations and Gene-Expression Regulations

by

Kayode A. Sosina

A dissertation submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
May, 2020

Abstract

This dissertation focuses on approaches which integrate both gene-expression data and genetic association results to extend insight into disease pathogenesis. We investigate and proffer strategies to accurately estimate cell type composition using both single cell and single nucleus RNA-Seq data in Brain (Chapter 2). We map cis regulatory effects in the retinal transcriptome and integrate data about trait associated variants to identify potential target genes which are involved in Age-related Macular Degeneration (Chapter 3). We propose a likelihood framework to model trans-regulatory effect size distribution and apply the method to four tissues, using summary level data, from the Genotype-Tissue Expression Consortium (Chapter 4).

Primary Readers

Nilanjan Chatterjee (Advisor)
Professor
Department of Biostatistics & Department of Oncology
Bloomberg School of Public Health, The Johns Hopkins University

Jeffrey Tullis Leek (co-Advisor)
Professor
Department of Biostatistics

Bloomberg School of Public Health, The Johns Hopkins University

Andrew Jaffe
Associate Professor
Department of Mental Health & Department of Biostatistics
School of Medicine, The Johns Hopkins University

Alexis Battle
Associate Professor
Department of Biomedical Engineering & Department of Computer Science
Whiting School of Engineering, The Johns Hopkins University

Alternate Readers

Priya Duggal
Associate Professor
Department of Epidemiology
Bloomberg School of Public Health, The Johns Hopkins University

Ni Zhao
Assistant Professor
Department of Biostatistics
Bloomberg School of Public Health, The Johns Hopkins University

Acknowledgments

First, and foremost, I would like to thank my adviser, Dr. Nilanjan Chatterjee, for his fundamental role in my doctoral work. Without his expertise, patience, guidance and support for the past four years, the bulk of this dissertation would not have been possible. You've been a great teacher and an incredible mentor.

I would like to thank Dr Jeffrey T. Leek, my co-advisor, and Dr Andrew E. Jaffe for their support, encouragement, and great advice. I offer my sincere gratitude to Dr Alexis Battle for her invaluable insights, and being a great collaborator. I gratefully acknowledge the members of my Ph.D. committee for their time and valuable feedback on a preliminary version of this doctoral work.

Table of Contents

List of Tables

List of Figures

B.8 **Differential expression and WGCNA analysis**. **a**, Heatmap showing the expression pattern of differentially expressed genes by comparing advanced AMD to controls with and without adjusting for age at the significance threshold at FDR$\leq$0.20 **b**, We identified 47 modules, each containing between 16 and 4,847 genes. Top: Dendrogram of genes with topological overlap used as distance (shown on y-axis). The color bar below indicates which module the genes belong to. Bottom: Hierarchical clustering of module expression eigenvalues (eigengenes). The modules involved in complement (yellow), angiogenesis (light green), immune activation (magenta), and extracellular matrix (pink) are highlighted in red. These modules were adjacent to each other according to eigenvalue-based hierarchical clustering. **c**, Two of these modules were particularly interesting as they were enriched for literature (pink FDR= 2.21×10^{-3}; magenta FDR= 1.37^{-9}) and leading edge (pink FDR = 1.10×10^{-3}; magenta FDR= 1.33×10^{-26}) candidate genes. Additionally, the magenta module was enriched for genes from the GWAS loci (FDR= 2.38×10^{-4}). The pink module also contained three DE- (FBLN1, MOXD1, IGFBP7) and two AMD-associated genes (COL8A1 and MMP19). GO analysis of the magenta and pink module highlighted extracellular matrix organization and immune response pathways, respectively, which were previously implicated in AMD pathology. These modules interacted closely with two other modules; the light green (also enriched for literature genes, FDR= 8.30×10^{-3}) and light yellow, which were enriched for angiogenesis and complement GO terms, respectively. We show only genes that fall in either literature, GWAS, or differentially expressed groups and are strongly correlated with another such candidates (adjacency$>$ 0.05). 140

Chapter 1

Introduction

Interest in the post-GWAS era has moved from discovery to gaining mechanistic insights about trait-associated variants. Since most associations lie in non-coding regions of the genome, hence are likely involved in the disease/trait process through gene regulation, attention has been placed on the integration of RNA-seq data with GWAs results. To identify variants involved in gene regulation, both cis (local) and trans(distal) regulatory effects of these variants on gene expression are been studied across diverse tissues[1, 2, 3, 4, 5, 6] with both approaches having varying degrees of success. Analyses concerning cis-regulation, in particular, have been more successful relative to trans-regulation, with cis-regulation being shown to be more tissue agnostic, have larger effects, and have been helpful in identifying likely causal genes that are involved in the disease process[2, 3, 7, 8]. In contrast, results from trans-regulation have been sparse, mostly because of their much smaller effect sizes. Despite this, current findings show that trans-regulatory effects are more tissue-dependent, thus

providing valuable insights about tissue-specific mechanisms in the disease pathogenesis. Furthermore, trans-effects have also be shown to extend insights for loci with multiple cis effects, used to reveal novel pathways for a given phenotype and identify relationships between trait-associated SNPs that are independent[2, 3].

This dissertation focuses on approaches that integrate both gene-expression and genotype data to provide mechanistic insights for genetic associations by means of gene-expression regulation. Gene expression levels are affected by both environmental and genetic effects. For instance, we know that expression levels can vary significantly between different tissues[1, 2], cell types[9, 10, 11], and even over time[12]. Furthermore, multiple studies[2, 3] have shown that gene regulatory effects can vary by cell type. These variations can be used to identify cis-regulatory effects, and aid in interpreting regulatory variants underlying complex disease risk. However, genomic profiles corresponding to these cell types are often unobserved since bulk RNA-Seq is generated based on tissues which consist of mixtures of cell types (i..e., cellular composition). Hence, multiple approaches have been developed to estimate (or deconvolve) the relative or absolute amounts of each cell type in a given tissue. In Chapter 1 we focus on such approaches and show that several existing deconvolution algorithms which estimate the RNA composition of homogenate tissue, relates to the amount of RNA attributable to each cell type, and not the cellular composition (absolute estimates) relating to the underlying fraction of cells. We show that incorporating "cell size" parameters into RNA-based deconvolution algorithms can successfully recover cellular fractions in homogenate brain RNA-seq data. Furthermore, we show that using both cell sizes and cell type-specific gene expression profiles

from brain regions other than the target/user-provided bulk tissue RNA-seq dataset consistently results in biased cell fractions.

In Chapter 2, we integrate data from retinal transcriptomes, covering 13,662 protein-coding and 1,462 non-coding genes, with genotypes at over 9 million common single nucleotide polymorphisms (SNPs) for expression quantitative trait loci (eQTL) analysis of a tissue not included in Genotype-Tissue Expression (GTEx) and other large datasets[13, 1, 2]. Furthermore, by combining findings from transcriptome-wide association analysis (TWAS), colocalization analysis, cis-eQTL analysis, and AMD GWAS we able to extend insights for trait-associated variants. In Chapter 3, we propose a likelihood framework, using summary level, to estimate both the heritability and polygenicity (the number of loci that contribute to heritability) of trans-regulatory effects. We circumvent issues related to small sample sizes that are seen in current studies involving trans-eQTLs by marginalizing across genes and have developed a computationally efficient approach to summarize the data. Hence, dealing with the large number of marginal effects typically encountered in trans-eQTL analyses. Subsequently, we applied our model, using summary level data, to four tissues from GTEx V8.

Chapter 2

Strategies for cellular deconvolution in human brain RNA sequencing data

2.1 Introduction

Homogenate tissues like brain and blood contain a mixture of cell types which can each have unique genomic profiles, and these mixtures of cell types, termed "cellular composition", can vary across samples[14]. The importance of considering cellular composition within heterogeneous tissue sources has been highlighted in epigenetics research over the past several years[14, 15, 16], as, generally, failure to account for cellular composition when analyzing heterogeneous tissue sources can increase both false positives and negatives[17]. Previous work has identified widespread epigenetic

differences between neurons and glia using DNA methylation (DNAm) data[16, 18], and false positives may arise when there are cellular composition differences associated with dissection variability, disease, normal development or any other outcome of interest. For example, loss of neurons (or glia) because of disease may cause spurious loci associations with illness that stem solely from differing cellular compositions between disease states, or cell-type specific biological differences may exist that become more difficult to detect in the presence of unaffected cell types.

Statistical algorithms estimate the relative or absolute amounts of each cell type in the homogenate tissue data. These so called "cellular deconvolution" algorithms have been especially popular using DNAm data[19] as DNAm levels are constrained between 0 and 1 and are binary within single cells (i.e., individual CpGs are either methylated or unmethylated). These deconvolution algorithms can be classified into two general types, termed "referenced-based" and "reference-free"[19, 20]. Reference-free approaches only require as input an estimate of the number of potential cell types in a particular dataset (which can be non-trivial), and return latent components that preferentially capture cellular heterogeneity that can be adjusted for in differential methylation analysis[14, 19, 21]. However, these approaches do not return fractions of cells and may capture potential batch effects in addition to cellular composition. Conversely, reference-based approaches require cell type-specific genomic profiles for each cell type of interest as an input and return the relative fraction of each input cell type for each queried bulk sample[15], akin to an in silico cell counter. This class of algorithms therefore requires the generation of potentially many pure cell populations, which are typically generated from flow cytometry for applications to

DNAm data from bulk tissue.

While DNAm data can generate accurate absolute cell fractions in homogenate brain tissue[16, 18, 22] , there are several important considerations limiting more widespread application. First, RNA and gene expression profiling has been much more popular in postmortem brain studies, with more samples profiled with RNA sequencing (RNA-seq) than DNAm microarrays or sequencing. Secondly, the two cell classes typically used by DNAm-deconvolution algorithms are likely too broad to identify more subtle differences in dissection variability and potential stereological differences[23, 24]. While recent work has extended the number of cell populations that can be isolated by antibodies to separate neurons into their excitatory and inhibitory subclasses and oligodendrocytes from other glia[25], there are likely very few additional cell types that are possible to isolate using nuclear antibodies for DNAm samples. Researchers have therefore turned to using cell type-specific RNA microarray and sequencing datasets to adapt these reference-based deconvolution algorithms to homogenate RNA-seq samples[10, 11, 20, 26, 27, 28, 29, 30, 31, 32, 33, 34]. The majority of these studies have focused on tissues other than the brain, which can be freshly obtained and dissociated into individual cells for single cell RNA-seq (scRNA-seq) or be sorted into specific cell populations using flow cytometry for cell type-specific expression profiling. For example, the popular CIBERSORT approach[11] was designed for blood gene expression microarray data, but has been adapted to RNA-seq datasets in other tissues. Several of the above algorithms have been designed, adapted or implemented for brain tissue, including linear regression

followed by quadratic programming using the Houseman algorithm[15, 33, 34], non-negative least squares[32], the support vector machine-based CIBERSORT[11], the empirical Bayes method MIND[30], and MuSiC, which combines a recursive tree based approach with weighted non-negative least squares for cell type proportion estimation[26].

However, few of these approaches have validated that the resulting composition estimates are accurate, i.e, are absolutely similar to the true underlying composition, particularly in brain tissue. No approach to our knowledge has quantified the consequences of parameter and algorithm choices when only non-ideal reference data is available (e.g., mismatched tissue type, species, sequencing protocol, etc.), which occurs in almost all applications. Many reference datasets have been constructed from purified cell type-specific RNA-seq data from mouse[35], or RNA-seq data from sorted or dissociated nuclei in humans[36, 37, 38, 39, 40], and not whole cells, which are typically profiled in homogenate sequencing studies. Gene expression levels are also quantitative within individual cells (and not binary like in DNAm data) and the necessity of absolute expression levels for absolute composition quantification has largely been overlooked.

Here we directly evaluated the absolute accuracy of several popular RNA-seq-based deconvolution strategies using several different reference datasets including a bulk/homogenate dataset with paired DNAm and RNA-seq data from the nucleus accumbens (NAc) from 200+ deceased individuals[41]. We used the DNAm data to estimate absolute neuronal fractions for each sample, and evaluated absolute RNA-based deconvolution accuracy across a variety of scenarios. We first evaluated the

effects of using deep single cell RNA-seq (scRNA-seq) from healthy fresh human tissue obtained from surgically resected temporal cortex [9]. This dataset likely produces the most comparable RNA-seq profiles to frozen bulk postmortem tissue, since whole cells were profiled, and 90% of RNA is cytosolic in the cortex[42]. However, this dataset was derived from cells in a cortical brain region. We next produced snRNA-seq data from postmortem human NAc to use as a reference dataset, which results in potentially less comparable nuclear reference profiles but comes from a more comparable brain region. We lastly used cyclic-ouroboros single-molecule fluorescence in situ hybridization (osmFISH) imaging data from the somatosensory cortex region in mouse [43] to derive important parameters in popular deconvolution algorithms. Together, our results demonstrate that many algorithms are not accurate, even when estimating only two cell classes (neurons and glia), and we offer several strategies to assess and improve accuracy that can be applied across multiple datasets and cell types.

2.2 Methods

2.2.1 Estimation procedures

HOUSEMAN. This algorithm (Houseman et al., 2012) uses a linearly constrained quadratic optimization approach with additional non-negative constraints on the parameters. The linear constraint does not require that the sum of all coefficients equal one. This allows the possibility of unknown cell types in case the specification is not comprehensive. It was implemented using the minfi R Bioconductor package[44].

MuSiC. The MuSiC (Wang et al., 2019) approach models the relationship between the relative abundance of gene g in the bulk RNA-seq data and the mean expression level of the same gene in the reference dataset for a given individual. The relationship is provided below

$$Y_g \propto \sum_{k=1}^{K} p^k S^k \theta_g^k$$

Where $k = 1, \ldots, K$ is the index of the cell types, p^k is the proportion of cells from cell type k, and θ_g^k is the relative abundance for the g^{th} gene with respect to the kth cell type. S^k is the cell size parameter and is defined as the average number of total mRNA molecules for cell type k. By default, S^k is estimated automatically by MuSiC. For the deconvolution method comparisons that assessed cell size impact on neuronal cell type proportion estimation, S^k was derived from one of multiple data sources (Table 2.1) using 1) default settings 2) osmFISH or 3) the average number of total mRNA molecules for cell type k using only the top 25 or 50 most discriminating genes per cell type. We defined "most discriminating" as genes with the smallest p-values and fold change > 0.25, relative to other cell types. All estimation was carried out using the MuSiC package in R.

CIBERSORT. CIBERSORT uses a machine learning approach called ν-support vector regression (Newman et al., 2015; Schlkopf et al., 2000) and requires at least 2 input datasets to work. The first is a signature matrix that identifies the set of genes that are informative for the deconvolution procedure. The second is a bulk RNA-seq dataset to estimate cell type proportions.

The signature matrix depends on the tissue of interest. We generated a custom signature matrix. Using the Darmanis reference dataset, we generated both a reference sample file (gene-by-cell matrix) and a phenotype classes file (cell type-by-dummy variable identifying the cell type for each cell) and used the default setting (https://cibersort.stanford.edu/) to obtain a custom signature gene expression matrix. The specified false discovery rate (FDR) threshold used to include genes in the signature matrix was 0.30 (i.e. q = 0.30, default). Using this signature matrix, we then performed deconvolution on our bulk NAc RNA-seq data. As suggested in the documentation for CIBERSORT (https://cibersort.stanford.edu/), we disabled quantile normalization for our RNA-seq data.

NNLS/MIND. This is a simple linear regression with non-negativity constraints on the parameter estimates. The estimated fractions are then the value of each parameter estimate divided by the sum of all parameter estimates across cell types. MIND (https://github.com/randel/MIND) uses NNLS to estimate cell type fractions.

2.2.2 Bulk NAc Data Generation and Processing

Data generation and processing were described extensively in Markunas et al. 2019[41] and in Appendix A.

2.2.3 Reference Datasets

Darmanis (Darmanis et al., 2015). scRNA-seq data for 58,037 genes and 556 cells were obtained for brain samples across 8 individuals, as described previously (Darmanis et al., 2015). We filtered this dataset by removing cells based on embryonic samples and retaining cells from one of the following five cell types; Neuronal, Oligodendrocyte progenitor cells (OPC), Astrocytes, Oligodendrocytes, and Microglia. We also removed genes that had no expression for all cells in the reference dataset or did not show any expression in the bulk dataset (i.e., mean and variance zero). In total, we used 265 cells for this reference and 24,048 genes to estimate the cell type proportions for the 223 samples with bulk NAc data.

Single-nucleus RNA-seq data generation and processing in nucleus accumbens. Details are provided in Appendix A

2.3 Results

We motivate this work with a large human postmortem brain genomic dataset from the NAc, a brain region containing functionally distinct cell types critical in reward-processing and addiction[45, 46]. Genomic data from this region has been underrepresented in postmortem human brain sequencing studies, which have primarily focused on the frontal cortex[32, 47, 48] but its underrepresentation allows us to more

comprehensively evaluate the accuracy of cellular deconvolution using potentially imperfect and/or mismatched reference datasets (described below). We dissected homogenate NAc tissue from the ventral striatum (anterior to the optic chiasm) across 223 adult donors and concurrently extracted DNA and RNA from the exact same tissue aliquot (see Methods), which allows for directly comparable cellular composition in each fraction. We profiled genome-wide DNAm with the Illumina Infinium MethylationEPIC microarray and performed reference-based deconvolution to estimate the fraction of neurons in each sample (see Methods). We have previously demonstrated the absolute accuracy of the Houseman deconvolution algorithm[15] in postmortem human brain DNAm data[18, 22]; here we found very high correlation (ρ = -0.949, Figure A.1) between the neuronal fraction and the first principal component (PC) of the entire DNAm profile (32.3% of variance explained), which we have shown to be an accurate surrogate of composition in frontal cortex[49] and blood[14]. The corresponding RNA was sequenced using the Illumina sequencing with RiboZero Gold library preparations (see Methods). This "gold standard" dataset, therefore, has DNAm-derived neuronal composition values and RNA-seq data from 223 samples to explore the accuracy and concordance of many popular cellular deconvolution algorithms.

2.3.1 Mismatched reference datasets bias deconvolution

We first assessed the accuracy and concordance of four reference-based deconvolution algorithms: Houseman, CIBERSORT, NNLS/MIND, and MuSiC for two cell populations - neurons and non-neurons/glia - in our NAc RNA-seq dataset using

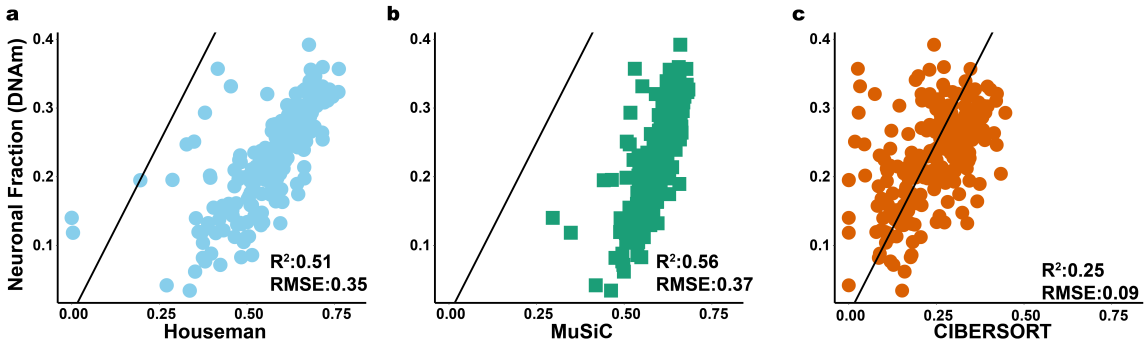


Fig. 2.1. Deconvolution in bulk NAc data using gene expression profiles from the temporal cortex (Darmanis); Scatter plots showing the estimated neuronal proportions across the 223 individuals using the Houseman approach for DNAm reference vs neuronal proportions estimated using (**a**) the Houseman approach with scRNA reference, (**b**) MuSiC with default settings and scRNA reference data, and (**c**) CIBERSORT with scRNA reference data.

recommended default settings (see Methods). We initially used single-cell RNA-seq (scRNA-seq) data from the temporal cortex of eight adult donors obtained during surgical resection generated and described in Darmanis et al., 2015[9] as the cell type-specific reference profiles for these algorithms. Importantly, these reference data were generated from fresh tissue, which preserved the integrity of the cells and corresponding cytosolic RNA, the predominant fraction of total RNA from brain[42] profiled in homogenate tissue. Furthermore, these reference profiles provide coverage of entire transcripts (as opposed to only the 3′ ends) using Fluidigm C1 sequencing. Therefore, these expression profiles should be more comparable to bulk brain sequencing studies, with the caveat that the reference dataset was obtained from a different brain region (temporal cortex versus NAc and from living subjects as opposed to postmortem subjects).

We used measures of root mean square error (RMSE) to assess accuracy and

squared Pearson correlation coefficients (R^2) to assess concordance for each algorithm's estimated neuronal fraction compared to the DNAm-based neuronal fractions (Figure 2.1). RMSE quantifies the degree of bias, i.e., how much our cell type estimates (RNA composition estimates) deviate from the absolute cell type fractions, with smaller values corresponding to the cellular composition and RNA composition being more similar. R^2 quantifies the amount of information our estimates contain about how the absolute cell type fractions vary in the population being studied, i.e., how much variability of the cell type fractions, across individuals, is captured by our composition estimates. Houseman (Figure 2.1a), MuSiC (Figure 2.1b), and NNLS produced concordant (high correlation; Houseman $R^2 = 0.51$, $p < 2.20 \times 10^{-16}$; MuSiC $R^2 = 0.56$, $p < 2.20 \times 10^{-16}$; NNLS $R^2 = 0.54$, $p < 2.20 \times 10^{-16}$) but biased (high RMSE, > 0.35) neuronal fraction estimates. CIBERSORT produced more discordant (moderate correlation; $R^2 = 0.25$, $p = 5.13 \times 10^{-03}$) neuronal fraction estimates (Figure 2.1c), but with less bias (low RSME, 0.09). We found that CIBERSORT, compared to either MuSiC or the Houseman RNA approach, was the most accurate. However, its estimates provided the least information (R^2 value) about the variability of the estimates based on DNAm data. In comparing the R^2 metric across the three approaches, we found that MuSiC provided the most information about the observed variability of the observed cell type proportions among the 223 individuals but was the most biased. These results suggest that all four of these approaches overestimated the proportion of neurons in bulk brain tissue, even under the simplest application to deconvoluting two distinct cell populations. However, it was unclear how much algorithm parameters and reference dataset differences (in regards to technology and

brain region) contributed to the performance of these methods.

2.3.2 Methods for reducing bias in cellular deconvolution

Table 2.1. Cell sizes used for deconvolution.

Cell type	NAc 50 genes (UMIs)	NAc 25 genes (UMIs)	NAc all genes (UMIs)	Temporal cortex, Darmanis et al. (Counts)	osmFISH cell Area (μm²)	osmFISH nRNA (intensity)
Glial	710.63	453.24	5763.55	12879.73	90.87	180.46
Neuronal	4513.58	2793.54	29884.65	18924.66	122.96	198.86
Neuronal /glial ratio	6.35	6.16	5.19	1.47	1.35	1.1

UMI, unique molecular identifier, counts= $\log_2$(cpm+0.5), intensity = sum of probe intensities across 24,048 genes.

Many of the above deconvolution strategies have several parameters whose adjustment could reduce the observed bias (i.e., maximize accuracy) and increase the concordance between these neuronal fractions. The MuSiC algorithm particularly has an interpretable "cell size" (see Methods) parameter used in the deconvolution process. Different cell types could have more or less absolute RNA abundance, for example if they were larger or smaller, or if they were more or less transcriptionally active. We hypothesized that the overestimation of neuronal fractions resulted from neurons being larger and more transcriptionally active. However, this "cell size" parameter, regularly defined by the algorithm as the average expression level for a given cell type summed across genes, is estimated directly from the reference cell

type-specific RNA-seq profiles by default (see Methods). However, some scRNA-seq (or snRNA-seq) library preparation and sequencing strategies, like the Fluidigm C1 system, may normalize cDNA libraries to the same concentration prior to sequencing, which will remove potential variability in RNA abundances across cell types. We therefore sought to use external data to better estimate these cell size parameters (Table 2.1) and assessed the resulting effects on cellular deconvolution accuracy.

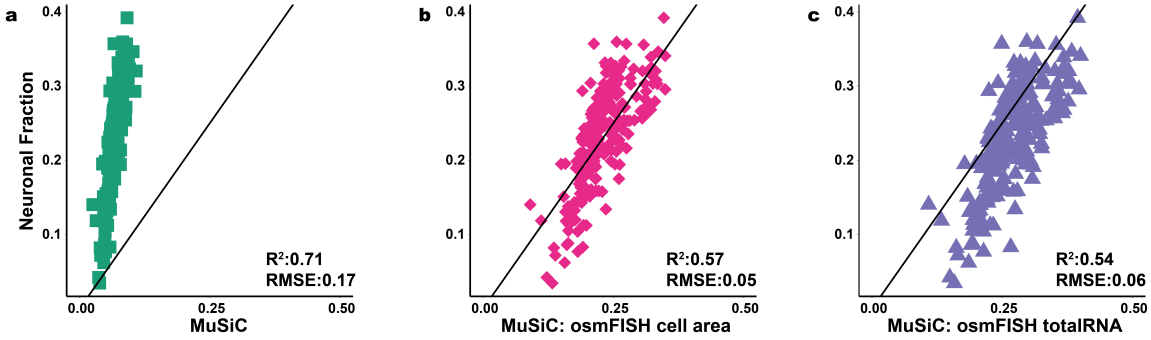


Fig. 2.2. Deconvolution in bulk NAc data based on a single nucleus RNAseq (snRNA-seq) reference dataset from the same brain region; Scatter plots comparing the estimated neuronal proportion obtained for each individual using the Houseman approach with DNAm reference dataset vs neuronal proportions obtained using (**a**) MuSiC with default settings and a snRNA-seq NAc reference dataset, (**b**) MuSiC based on a snRNA-seq NAc reference dataset with cell sizes for each cell type estimated using osmFISH cell area (mouse), and (**c**) MuSiC based on a snRNA-seq NAc reference dataset with cell sizes for each cell type estimated using osmFISH total RNA abundance (mouse) per cell type.

First, we used external ouroboros single-molecule fluorescence in situ hybridization (osmFISH) data from mouse somatosensory cortex[43] to construct two different types of cell size parameters for the MuSiC algorithm (as data from NAc did not exist). We extracted the estimates of both cell size (via their provided segmentations) and total RNA abundance (via the sum of all gene fluorescence signal) aggregated across neuronal and non-neuronal cell types. We subsequently utilized these estimates as proxies for cell size in human RNA-seq data when deconvoluting neuronal

fractions. In these data, comparing neurons to non-neurons, neurons were both larger (123 vs 91 μm^2, $p < 2.20 \times 10^{-16}$) and had more total RNA (199 vs 180 intensity, $p = 1.73 \times 10^{-05}$) as we observed in the estimated cell size in the MuSiC algorithm using the Darmanis dataset (18,925 vs 12,880 normalized counts). We did not observe any improvement in the concordance (osmFISH cell area $R^2 = 0.55$, $p < 2.2 \times 10^{-16}$; osmFISH totalRNA $R^2 = 0.54$, $p < 2.2 \times 10^{-16}$) or accuracy (osmFISH cell area RMSE = 0.39; osmFISH totalRNA RMSE = 0.43) of the estimated cell type fractions when we compared our results from default settings to those based on applying cell size proxies using mouse data (Figure A.2 a-b). These results may not be particularly surprising, given the numerous differences between mouse and human morphology, and the different brain regions profiled.

We then generated snRNA-seq dataset from 2 postmortem NAc donors and 4,169 total nuclei to produce more comparable cell type specific cell size (see Methods) parameters and reference expression profiles (see Methods). First, we used the NAc reference dataset at the single nucleus level and ran the MuSiC algorithm with default settings, which used both NAc-based cell sizes and expression profiles, to deconvolute neuronal fractions (Figure 2.2a). We confirmed that, on average, neurons had more total RNA than non-neurons using this NAc snRNA-seq dataset (103 vs. 72 unique molecular identifiers [UMIs] per gene, $p < 2.2 \times 10^{-16}$). Furthermore, while there was a high correlation among neuron-specific gene expression effects across the NAc and temporal cortex (Darmanis et al.) reference profile datasets, we observed genes with different magnitudes of effects based on differential expression results between neuronal and non-neuronal cell types (Figure A.33). When using

Table 2.2. **Bias and concordance results for deconvolution of bulk NAc data using each cell size and gene expression reference dataset.**

Method	Cell size	Reference dataset			
		scRNA-seq in temporal cortex (Darmanis et al.)		snRNA-seq in NAc	
		Concordance (R^2)	Accuracy (RMSE)	Concordance (R^2)	Accuracy (RMSE)
MuSiC	None	0.54	0.45	0.54	0.08
	scRNAseq in temporal cortex (Darmanis et al.)	0.56	0.37	0.58	0.05
	snRNAseq in NAc	0.59	0.08	0.71	0.17
	snRNAseq in NAc (Top 25 genes)	0.59	0.06	0.71	0.18
	snRNAseq in NAc (Top 50 genes)	0.59	0.05	0.72	0.18
	osmFISh-Cell area	0.55	0.38	0.57	0.05
	osmFISh-Total RNA	0.54	0.43	0.54	0.06
CIBERSORT	N/A	0.25	0.09	N/A	N/A
NNLS	N/A	0.54	0.40	0.72	0.22

both the NAc-based cell size and gene expression reference profiles, we observed a substantial improvement in both the concordance and the RMSE for the estimated neuronal fractions compared to using the temporal cortex dataset only. However, the estimates were still biased, and this bias increased as the neuronal fraction across individuals increased, suggesting that the NAc-based cell sizes together with the estimated abundance may be incorrectly characterizing the true underlying neuronal expression level for these individuals. Eliminating the cell size parameter resulted in similarly reduced concordancies in both the temporal cortex and the NAc reference datasets, but increased accuracy only using the NAc reference dataset (Table 2.2). This implies that the underlying broad cellular composition was well captured by the gene abundance information for a matched brain region.

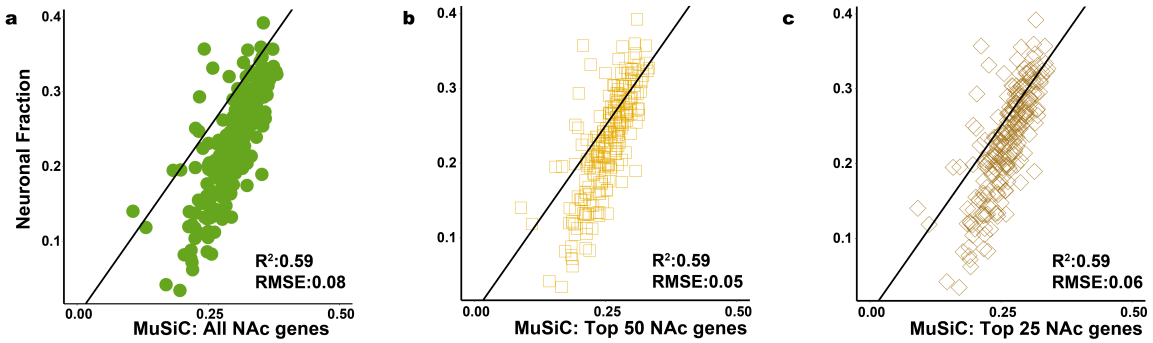


Fig. 2.3. Deconvolution in bulk NAc data using gene expression profiles from the temporal cortex (Darmanis et al) and different estimates of cell size; Scatter plots comparing the neuronal fraction estimated for each individual using DNAm data and the Houseman method vs neuronal fractions based on scRNA-seq data and estimated using MuSiC with (**a**) cell-size estimated using all genes expressed in the NAc snRNAseq reference dataset, (**b**) cell-size estimated using the top 50 cell type discriminating genes in the NAc snRNAseq reference dataset, and (**c**) cell-size estimated using the top 25 cell type discriminating genes in the NAc snRNAseq reference dataset.

We then combined different estimates of cell size parameters (NAc snRNA-seq

versus osmFISH) and gene expression reference profiles (NAc snRNA-seq versus temporal cortex scRNA-seq) and assessed the effects on deconvolution accuracy in bulk NAc RNA-seq data. When running MuSiC using the estimates of cell size based on osmFISH data with the NAc expression reference profiles, we observed further improvements in the bias of the estimated cell type fractions but saw a minimal difference in the concordance (Figure 2.2b and c). Surprisingly, when we used only the Darmanis cell type-specific expression levels, the best (least biased and most concordant) deconvolution results were produced using cell sizes estimated from NAc snRNA-seq data, with improvements in both the concordance and the RMSE (Figure 2.3a). Specifically, when we compared the R^2 and the RMSE estimates to those observed under the default setting for the Darmanis reference with the mismatched brain region, we see a small (6% relative change, $p = 3.3 \times 10^{-02}$) increase for the concordance and a substantial (78% relative change, $p < 1 \times 10^{-04}$) decrease for the RMSE. We further refined the NAc cell size estimates using sets of the top 25 and 50 cell type discriminating genes (see Methods), which slightly improved our estimates of the absolute cell type fractions (Figure 2.3b and 2.3c). Both the concordance (7% relative change, $p = 3.5 \times 10^{-02}$) and RMSE (86% relative change, $p < 1 \times 10^{-04}$) improved even more when compared to the default approach using a mismatched reference dataset. Across all approaches, the most accurate (least biased) result occurred when we used cell sizes estimated from Darmanis scRNAseq data and gene expression from NAc snRNA-seq data, while the most concordant results were observed when we used NAc snRNAseq data exclusively.

In summary, when we used a region-matched appropriate dataset - NAc snRNA

data - as the reference, or to derive estimates of the cell size, we observed that estimates of the cell type proportions generally improved (Table 2.2, Figure 2.2 a and Figure 2.3a-c). In settings where we had a mismatched reference dataset (e.g., mismatched on brain region or species), incorporating estimated cell sizes obtained from the matched brain region (NAc) provided the best result in metrics for both concordance and accuracy, and we slightly improved these metrics when we refined the gene sets used to estimate the cell sizes.

2.4 Discussion

Statistical deconvolution strategies have emerged over the past decade to estimate the proportion of various cell populations in homogenate tissue sources like blood and brain from both gene expression and DNAm data. Our results together suggest that many existing RNA deconvolution algorithms estimate the RNA composition of homogenate tissue, e.g. the amount of RNA attributable to each cell type, and not the cellular composition, which relates to the underlying fraction of cells. This was evident by the consistent overestimation of larger and more transcriptionally active neuronal cells. We have identified that incorporating cell size parameters into RNA-based deconvolution algorithms can successfully recover cellular fractions in homogenate brain RNA-seq data. We have lastly shown that using both cell sizes and cell type-specific gene expression profiles from brain regions other than the target/user-provided bulk tissue RNA-seq dataset consistently resulted in overestimating neuronal fractions. We have developed an extension of the MuSiC framework

[26] that allows for the incorporation of independent cell size estimates, and have further provided cell size estimates for human brain (shown in Table 2.1) as a part of the package: https://github.com/xuranw/MuSiC.

Characterizing cellular heterogeneity is especially important in human brain, where the underlying cell types can have diverse functions and disease associations that could be missed in studies of bulk tissue [17]. Here we show that RNA-based deconvolution for just two cell populations - neurons and non-neurons - largely fails to estimate the underlying cellular composition of bulk human brain tissue across a variety of algorithms and strategies. We quantified the diverse range of neuronal fractions estimated by several popular algorithms to better understand the effects of reference cell type-specific expression profiles and differences in cell size and/or activity profiles on deconvolution. We specifically examined the common scenario of performing RNA deconvolution using cell type-specific reference datasets that can be fundamentally different from user-provided homogenate tissue target datasets, for example differing in profiled brain region, sequencing technology and/or cellular compartment. These problems are likely magnified in human brain tissue compared to suspended cells like blood, where deconvolution strategies are more easily validated against true cell fractions obtained by routine complete cell counts [11]. We lastly emphasize caution when performing RNA-based deconvolution using many cell types (i.e., more finely-partitioned cell classes) without having the ability to validate cell counts on at least a subset of samples.

We therefore offer several recommendations for performing RNA-based deconvolution in bulk human brain gene expression data, particularly when aiming to identify

cellular, and not RNA, composition.

1. Providing estimates of cell size for each reference cell type improves the concordance and reduces bias when performing RNA deconvolution to estimate cellular fractions. Biologically-motivated and valid external estimates of cell-size improve the accuracy of the estimated cell type fractions, even when gene expression profiles for reference cell populations are obtained from other brain regions (Figure 2.3). The exact biological interpretation of these estimated cell sizes, particularly when estimated across species, is arguably unclear, but likely relates to correcting for absolute RNA abundance and differences in transcriptional activity between cell populations. Regardless of the method used for deriving cell sizes, neurons consistently had more RNA than glia. We note that our recommended strategies for estimating cell size have only been assessed for broad classes of cell types, and further work is needed to validate extensions to more stratified subclasses of cells.

2. The concordance and bias improvements using full-length single cell sequencing from a different brain region (temporal cortex), rather than single nuclei RNA-seq from the target brain region (NAc) highlighted the importance of comparability between reference gene expression profiles and the homogenate tissue expression levels. While previous reports have identified high correlation between nuclear and cytosolic gene expression levels in both bulk[50] and single cell[39, 51] resolution, comparable absolute (and not relative) expression levels are seemingly important for the accuracy of these RNA-based cellular deconvolution algorithms. There further is an experimental design tradeoff between

profiling more nuclei (1000s) using 3′ technologies like 10x Genomics Chromium Single Cell Gene Expression compared to profiling fewer nuclei (or cells, 100s) using full-length sequencing technologies like SMART-seq if researchers wish to generate their own reference profiles.

3. Using reference cell type-specific expression profiles from comparable brain regions as the bulk RNA-seq target dataset is important, and can especially greatly increase the concordance of these RNA deconvolution strategies with neuronal fractions.

The choice of maximizing accuracy (by minimizing bias) versus increasing concordance in assessing these algorithms is an important consideration, particularly when generating custom expression reference profiles is prohibitive (Table 2.2). These two objectives largely relate to whether the goal of RNA deconvolution is to estimate cell fractions (and maximize accuracy) or RNA fractions (and maximize concordance). Estimation of RNA fractions (by maximizing concordance) may be sufficient to control for potential confounding due to composition differences between outcome groups [17]. We note this can also be accomplished using "reference-free" deconvolution [19] or through the estimation of potentially sparse principal components[14, 21] that control for relative differences in cellular composition. However, estimation of cellular fractions (and maximizing accuracy) is arguably more useful, both for assessing human brain tissue dissection during data generation and to identify cell type-specific effects when using these cellular fractions in downstream differential expression analyses[52].

Together, our results demonstrate that many RNA deconvolution algorithms do

not produce accurate cellular fractions when estimating only two cell classes (neurons and non-neurons). We offer several strategies and corresponding software to assess and improve accuracy that can be applied across multiple datasets and cell types.

Chapter 3

Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration

3.1 Introduction

AMD, a leading cause of incurable vision impairment, results in progressive loss of photoreceptors, particularly in the macular region of the retina[53]. AMD-GWAS have identified strong and highly replicated association of 52 independent SNPs at 34 genetic loci accounting for more than 50% of the heritability[54]. To derive mechanistic insights and further advance AMD genetics, we initiated the EyeGEx project to

elucidate genetic regulation of gene expression in the human retina. We characterized 523 postmortem retinas from 517 donors by using the Minnesota Grading System (MGS)[55], with criteria similar to the Age-related Eye Disease Study (AREDS)[56] (Fig. B.1 and Supplementary Data 1). MGS1 donor retinas demonstrated no AMD features and served as controls, whereas MGS2 to MGS4 samples represented progressively more severe disease stages.

3.2 Methods

3.2.1 Study subjects

Postmortem human donor eyes were procured by the Minnesota Lions Eye Bank after informed consent from the donor or next of kin was obtained, in accordance with the tenets of the Declaration of Helsinki. These studies were approved by the institutional review boards of the University of Minnesota and National Eye Institute, National Institutes of Health. Exclusion criteria for donors included a history of diabetes or glaucoma. Donors were also excluded from this study if, upon examination of donor macular images, there were clinical symptoms of diabetic retinopathy, advanced glaucoma, myopic degeneration, or the presence of atypical debris in the eyes. Donor eyes were enucleated within 4h of death and stored in a moist chamber at 4°C until retinal dissection was performed. Dissection and classification of donor retinas for AMD were carried out according to the four-step MGS as previously described[55, 57]. Tissue sections were flash frozen in liquid nitrogen and stored at

-80 C until further processing. Samples with ambiguous or no MGS levels were excluded from downstream analysis. Details of donor characteristics are described in Appendix B.

3.2.2 RNA-seq, genotyping, and quality control

Details of RNA-seq, genotyping, and quality control are provided in Appendix B.

Batch correction

Surrogate variables were identified and estimated for known batch effects as well as latent factors by using the supervised SVA (SSVA version 3.28.0/3.24.4) method[58, 59, 60] based on the following model:

$$\text{gene expression} \sim \text{MGS} + \text{Sex} + \text{Age}$$

Negative-control genes for SSVA were selected from a reported list of 3,804 housekeeping genes that are uniformly expressed across 16 human tissues[61]. The Pearson method was used to determine correlations between all significant surrogate variables identified by SSVA and possible sources of variation, including biological and technical factors. Known batch effects were assessed with Principal Variance Component Analysis (PVCA) (version 1.23.0) before and after batch correction[62]. All surrogate variables identified by SSVA were used for batch correction. Additional details are described in Appendix B.

3.2.3 Reference transcriptome

We generated the control human retina transcriptome profile from 105 MGS1 retinas by applying two criteria for gene expression: the first was to remove weakly expressed genes across all MGS stages (i.e., $\geq$1 CPM in $\geq$10% of all 453 samples), and the second was to describe the transcriptomic landscape in the retina with greater confidence (i.e., $\geq$2 CPM in $\geq$50% of all 105 MGS1 samples). We calculated the cumulative transcriptional output as previously defined[63] by converting CPM into FPKM values to take gene length into account. Similarities in transcriptomes between the retina and 53 GTEx tissues were observed with a gene filter of $\geq$1 CPM in $\geq$10% of all samples across all tissues, whereas a different gene filter, namely $\geq$1 CPM in $\geq$10% of samples within each tissue, was applied to identify genes whose expression was at least tenfold higher in the retina than in other tissues. Pathway enrichment analysis was performed with GO biological-process terms[64, 65] within clusterProfiler version 3.4.4 (ref. [66]) by using a Benjamini-Hochberg-adjusted P value $\leq$ 0.05 as the significance threshold. The analysis and classification of potentially novel isoforms of known genes and unknown intergenic transcripts were performed with the Cufflinks suite, version 2.21[67, 68]. Further details are provided in Appendix B.

3.2.4 cis-eQTL mapping

The analysis included 406 individuals for whom genotype and retina gene expression data were available, 17,389 genes expressed at $\geq$1 CPM in at least 10% of the retina samples, and 8,924,684 genotyped and imputed common variants. Cis-eQTL analysis was conducted with QTLtools version 1.0[69], with a linear model to adjust for

disease status (MGS level), age, sex, population stratification (ten principal components), and batch effects (21 surrogate variables). In the first step of the analysis, the variant most associated with each gene was selected, and then permutation was used to determine the distribution of its test statistic under the null. This procedure was subsequently used to obtain the P value for each gene. These P values were adjusted for multiple testing with the q-value approach[70] at the desired type I-error level. The second step of the analysis involved the identification of all eVariants with independent effects on a given eGene (significant gene from the first stage). This step was done by using the gene-level thresholds derived from the first stage and then identifying which variants exhibited nominal P values below these thresholds, on the basis of the forward-backward stepwise regression algorithm.

3.2.5 GTEx comparison

To calculate π_1, we compared our cis-eQTL discoveries by using the following definition:

$\pi_1 = P$(cis-eQTL in discovery tissue is significant in replication tissue|cis-eQTL in discovery tissue was also analyzed in the replication tissue)

Thus, for each cis-eQTL (gene-variant combination) we required that the combination be analyzed in both tissues being compared.

3.2.6 GWAS lead-variant analysis

Forty-one lead variants from AMD-GWAS3 were analyzed. Those not found either were not in the reference dataset used for imputation (six variants) or did not pass our MAF threshold of $< 1\%$ (five variants). Matrix eQTL version 2.1.1[71] was then used to obtain the marginal associations by using the same cis criteria, which were then corrected for multiple testing only for the number of variants tested, by using the Bonferroni method with a type I error rate of 5%.

3.2.7 Enrichment

In general, we processed quantile-quantile plots for each GWAS dataset by removing all SNPs within ± 1 Mb of the known GWAS signals and subsetting to variants with a MAF of at least 5%, after removing variants in the major histocompatibility region. The remaining variants were then grouped according to eQTL characteristics. Details can be found in Appendix B.

3.2.8 Colocalization

Likely colocalizing variants between the eQTL and the GWAS data were identified with eCAVIAR version 2.0 (ref. [8]) (Appendix B) on the basis of marginal statistics from the cis-eQTL analysis and from AMD GWAS[54].

3.2.9 TWAS

To perform the TWAS, the log-transformed, SSVA-corrected expression data from the 406 samples in our dataset that both passed RNA-seq and genotyping quality control were inverse-normal transformed (rank offset=3/8)[72] to moderate the influence of potential outliers. Expression was then controlled for sex, age, and the ten population-structure variables determined by Eigenstrat version 7.2.1[70, 73]. For each gene, we took the subset of SNPs within 1 Mb of its start or end site that had GWAS statistics[54] by using VCFtools version 0.1.15[74]. TWAS implementation was performed according to Gusev et al.[75], heritability was calculated with GCTA version 1.21[76], and genetic control of expression was modeled with mixed models, LASSO, or elastic net (α=0.5), depending on which of the three methods produced the highest fivefold cross-validation R^2.

The effect sizes from these models acted as weights. Weighted z scores were summed for each gene, and this gene-trait association statistic was divided by its standard deviation while LD was accounted for between GWAS statistics. Standardized gene-level scores were tested against the standard normal distribution on both sides. The FDR was calculated to account for multiple testing across genes with calculated P values; genes that had an FDR<0.05 were considered significant. We also determined whether genes passed a 0.05 significance threshold after Bonferroni correction. Genes were then filtered according to their model expression fit; genes with a genetic model $R^2 < 0.01$ were discarded.

We also performed a permutation test to determine the role that the eQTL data

played in the associations: for genes with a TWAS P value$<$ 0.001, weights were randomly assigned to SNPs, and the gene-level z scores were recomputed for an adaptive number of iterations to generate a null distribution against which the original TWAS statistic was tested[75]. Details on the methods used for the conditional TWAS test can be found in Appendix B.

3.2.10 Differential expression

Differential expression was assessed with the limma package in R version 3.34.2[77] with a significance threshold of FDR$\leq$ 0.20. MGS was treated as an ordinal variable in pairwise comparisons between controls and each AMD stage. Differential expression was performed with adjustments for sex and batch effects (22 surrogate variables), with or without age as a covariate. Age was the most significant non-genetic risk factor for AMD, and age-related gene expression changes would probably be relevant to AMD. We therefore also performed differential expression analysis without correcting for age to generate a comprehensive list of candidate genes that require further investigation to ascertain their contribution to AMD pathogenesis. Additional differential expression analyses, performed after the removal of samples with conditions such as hypertension, high cholesterol, and cardiovascular disease, were consistent across all comparisons made (data not shown).

3.2.11 Gene set enrichment analysis and leading-edge analysis.

GSEA was performed by preranking genes by significance and the direction of fold change from differential expression analysis, and then testing for association with the GO biological-process gene set deposited in the GSEA MSigDB resource version 2.2.4[78]. Leading-edge analysis was performed on gene sets reaching a significance threshold of FDR$\leq$ 0.25 and absolute normalized enrichment score of $\geq$ 2.0. Significant gene sets were further classified into common functional categories by visualization of the GO structure as described in Appendix B (see URLs).

3.2.12 Weighted gene-correlation network analysis

Weighted gene-correlation network analysis[79] was performed on all 453 samples that passed RNA-seq quality control, to group genes by expression profile, with the associated software WGCNA version 1.51. The log-transformed expression values were corrected for age, sex, and batch effects (determined by SSVA[59, 60]). Adjacency was calculated with Spearman correlation, and the power by which we raised the absolute values of the correlation to obtain the adjacency matrix was k=3. Through hypergeometric testing, at a significance threshold of 0.05 after Bonferroni correction for multiple testing, modules were assessed for the enrichment of the following types of genes: (i) genes deemed relevant to macular-degeneration pathogenesis in the literature, (i) genes within 500 kb of the 34 AMD loci identified through GWAS[54], and (iii) genes identified as leading edge by GSEA[78]. A list of genes relevant to

AMD was obtained from a previous published study[80] and was updated through extensive PubMed searching (through December 2017) with one of several search terms (Appendix B). Pathway analysis was performed on each module with GO biological-process terms[64, 65] through clusterProfiler version 3.4.4[66]. The connections between genes in modules were visualized with Cytoscape version 3.5.1[81]

3.3 Results

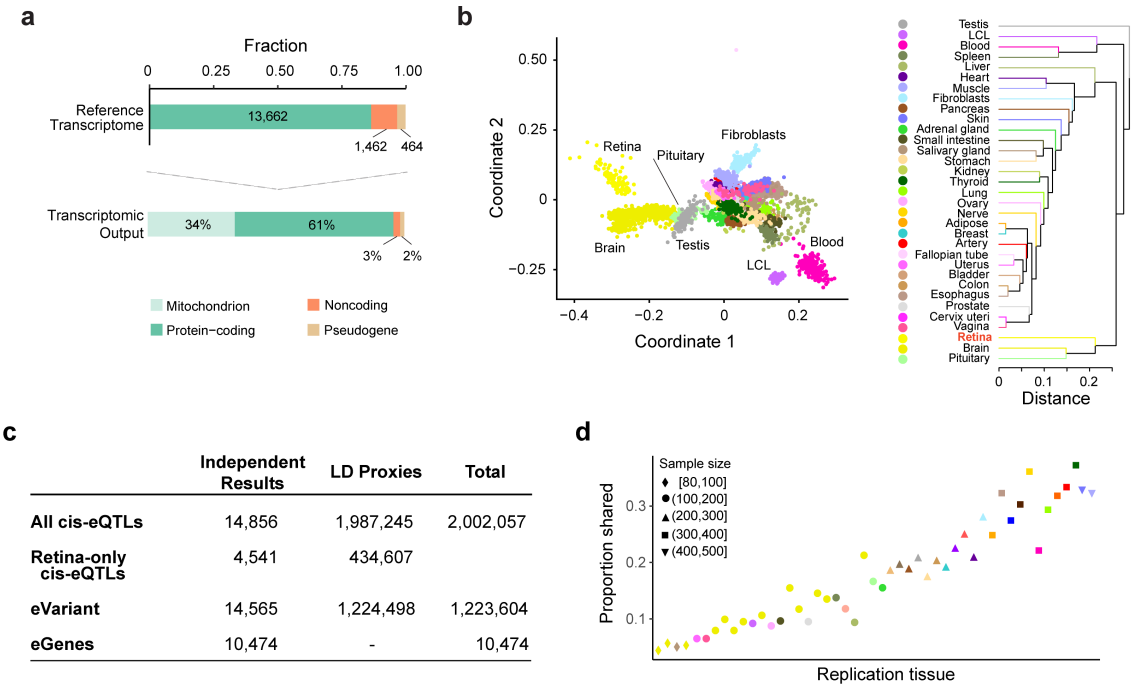
3.3.1 Samples and sequencing

RNA-seq of the donor retinas provided 32.5 million (median) uniquely mapped paired-end reads per sample, with a 94% mapping rate to Ensembl release GRCh38.p7 (Fig. B.2). After RNA-seq quality control (Appendix B), 105 MGS1, 175 MGS2, 112 MGS3, and 61 MGS4 samples were selected for further analyses. The reference-transcriptome profile was generated from MGS1 control retinas (Fig. 3.1a and Supplementary Data 2[7]) and included 67% of the protein-coding genes (13,662) and 6.7% of the noncoding genes (1,462) in Ensembl, in agreement with findings from a previous study[82]. High-abundance genes (186 genes showing$\geq$100 fragments per kilobase of transcript per million mapped reads (FPKM)) accounted for half of the Ensembl-annotated transcripts in our RNA-seq data and were enriched in visual perception, metabolic processes, and energy homeostasis (Fig. B.3a and Supplementary Data 2[7]). Overall, 34% of the retinal transcripts were of mitochondrial origin (Fig. 3.1a and Fig. B.3b), thus reflecting the high concentration of mitochondria in photoreceptors[83], the predominant cell type in the human retina[84].

Genome-guided transcript assembly supplemented 410 putative novel long intergenic noncoding RNAs (lincRNAs) and 2,861 proteincoding isoforms of genes expressed in the retina (Fig. B.3c and Supplementary Data 2[7]). The putative lincRNA isoforms were not enriched in any biological pathway. In contrast, predicted gene function and classification of novel protein-coding isoforms showed enrichment in Gene Ontology (GO) biological processes involving synapse structure or activity (adjusted P value=1.37×10^{-2}), sensory perception (adjusted P value=1.64×10^{-2}), regulation of membrane potential (adjusted P value=3.45×10^{-2}), and photoreceptor maintenance (adjusted P value=3.45×10^{-2}). The multidimensional scaling plot of the retina reference transcriptome against the GTEx v7 data distinguished tissue-specific clusters consistent with the defined biological replicates, whereas tissue hierarchical clustering on the mean gene expression levels revealed a high degree of similarity between the brain and retina (Fig. 3.1b, Fig. B.3d and Fig. B.4). We identified 247 genes with tenfold or higher expression in the retina than in at least 42 of the 53 GTEx (v7) tissues (Supplementary Data 2[7]).

3.3.2 A comprehensive resource of Retina eQTL

Mapping of cis-eQTLs (as defined by SNP-gene combination within ± 1 Mb of the transcriptional start site of each gene; Methods) identified 14,565 genetic variants (eVariants) controlling expression of 10,474 genes (eGenes) at a false-discovery rate (FDR) ≤ 0.05; these included 8,529 known protein-coding and 1,358 noncoding genes (Fig. 3.1c and Supplementary Data 3[7]). The strength of association was contingent on the eVariants distance from the transcriptional start site of its corresponding



	Independent Results	LD Proxies	Total
All cis-eQTLs	14,856	1,987,245	2,002,057
Retina-only cis-eQTLs	4,541	434,607	
eVariant	14,565	1,224,498	1,223,604
eGenes	10,474	-	10,474

Fig. 3.1. EyeGEx: Retinal transcriptome and eQTL analyses. **a**, Reference transcriptome output. Top: Fraction of expressed genes in Ensembl gene biotypes. Below: Percentage of gene expression in distinct gene subtypes. **b**, Within-tissue sample similarity and transcriptome comparison across the retina and the GTEx tissues (v7) based on normalized gene expression levels. Each color represents a distinct tissue. Left: multidimensional scaling. Right: tissue hierarchical clustering. **c**, A summary of retinal cis-eQTLs, eGenes and eVariants. 1.8% of the top eVariants (14,565) regulate more than one eGene. Variants in LD with the most significant eVariant are indicated as LD proxies. LD, linkage disequilibrium. **d**, The proportion of cis-eQTLs in the retina (y-axis) that are detected in GTEx (x-axis), ordered by the sample size of each tissue. Color and shape of each point represent the tissue and sample size, respectively.

eGene (Fig. B.5). Most of the retinal cis-eQTLs were present in at least one GTEx tissue, and more retinal eQTLs replicated with an increase in GTEx tissue sample size (Fig. 3.1d). The proportion of GTEx cis-eQTLs replicated in the retina was larger for GTEx tissues with smaller sample sizes[1](Fig. B.5f). Almost one-third of the retina-only eQTLs observed in our study, compared with those reported by GTEx for other tissues, were attributable to the relatively larger sample size (Fig.

B.6a,b).

3.3.3 Integrative analysis of Retina cis-regulatory effects with AMD risk

We examined the global role of eQTLs in the genetics of AMD. Q-Q plots identified cis-eQTL SNPs to be enriched for AMD associations with more pronounced enrichment for eVariants shared across several tissues[85, 86], and this relationship remained relatively consistent across all other complex disease phenotypes examined (see Fig. B.5g). We then integrated retina eQTL results with associations reported across loci identified by AMD-GWAS (Table B.1). Nine lead SNPs at the GWAS loci were significant eQTLs in the retina for 19 SNP-gene associations. Similar analysis showed a comparable number of lead SNPs as eQTLs in several GTEx tissues (for details see Supplementary Data 3[7]). To ascertain the most likely causal variants, we applied eCAVIAR which calculates the colocalization posterior probability (CLPP) to identify the variant responsible for both AMD-GWAS and retina-eQTL signals, after accounting for local linkage disequilibrium (LD) patterns. At the recommended threshold of 1% CLPP[8], we discovered likely causal SNPs and underlying target genes at six AMD loci (Table B.1, Fig. 3.2a). At two of these loci (B3GALTL and RDH5/CD63), the lead GWAS signal was identified as the most likely causal SNP, whereas the likely causal variant was distinct from the lead SNP at four other loci; SLC16A8 (rs5756908), ACAD10 (rs7398705), TMEM/VTN (rs241777), and APOE (rs157580) (Table B.1).

We then leveraged retinal eQTLs and the most recent GWAS data 3 to detect

novel AMD risk genes in a transcriptome-wide association study (TWAS) 14 using our retina transcriptome data. Gene expression was modeled using SNPs within a 1 Mb window using mixed models, Least Absolute Shrinkage and Selection Operator (LASSO), and elastic net. The TWAS identified 61 transcriptome-wide significant gene-AMD associations (FDR$\leq$0.05), which passed a gene expression model fit filter (R^2>0.01) (Supplementary Data 4[7]). We detected 38 genes within 1 Mb of 13 AMD-GWAS loci, and of these, 28 passed genome-wide Bonferroni correction (Fig. 3.2b). TWAS analysis also revealed 23 genes outside the GWAS loci (Fig. 3.2c); these genes fell within 16 separate regions (+/− 1 Mb). Three of these - RLBP1, PARP12 and HIC1 - were the only significant genes in the region and remained so even after Bonferroni correction, thus representing the strongest new candidate AMD-associated genes (Fig. 3.2d). Conditional testing of the full 61 significant (FDR$\leq$0.05) candidates revealed 47 independent signals (α=0.05). A permutation test (see Methods) demonstrated that two of the genes (MTMR10 and SH3BGR), which were at least 1 Mb outside of any GWAS region, had TWAS associations significantly informed by eQTL data after Bonferroni correction for the number of genes permuted (α=0.05; Supplementary Data 4[7]). However, we note that the test is overly conservative in the presence of LD.

We compared the data from eQTL, eCAVIAR, and TWAS to highlight the most plausible target genes; B3GLCT and BLOC1S1 were each identified as the only target gene at two AMD loci by all three methods, whereas SH2B3, PLA2G12A, PILRB, and POLDIP2/TMEM199 were likely targets at four additional loci identified by two methods (Table 3.1 and Fig. B.7). A comparison of these findings with those reported

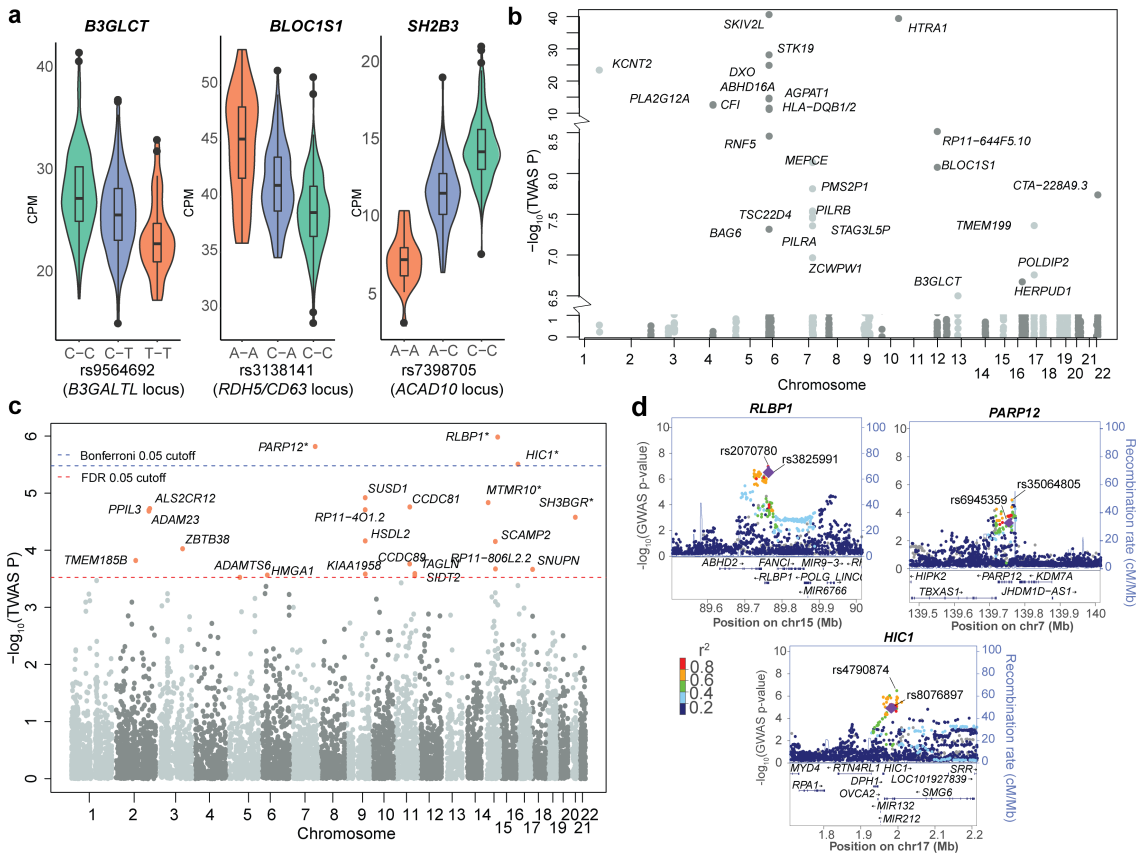


Fig. 3.2. Genes and variants associated with AMD. **a**, Violin plots of the relationship between the variant at a GWAS locus and the target gene identified by eCAVIAR. For three GWAS loci, the target gene (shown here) was the only one significantly associated (FDR ≤ 0.05) by TWAS. **b**, TWAS results for genes that pass Bonferroni-corrected significance identified within 1 Mb on either side of the lead SNP at previously-reported GWAS loci. PLEKHA1 (TWAS p-value $=7.91 \times 10^{-119}$) was omitted for appropriate scaling. **c**, Manhattan plot of TWAS-identified genes outside the reported lead SNP ($>$ 1 Mb on either side) at the GWAS loci. Of the genes with expression model $R^2 > 0.01$, 23 genes met the FDR threshold of 0.05 (red line), and three of these passed Bonferroni-corrected significance (cutoff shown as blue line). **d**, LocusZoom plots showing empirical GWAS association for top three TWAS signals outside GWAS loci. The diamonds indicate top eVariants for independent eQTL signals. The coloration of the points is determined by their LD with respect to the eQTL in purple. The top GWAS variant in the region is also labeled. The recombination rate is shown as a blue line.

in GTEx[1, 87] showed that the contribution of these SNPs to gene regulation varied across different tissues (Supplementary Data 3[7] and Appendix B.3.4). Specifically, no single nonretina tissue showed replication of the retinal findings for all SNP-target gene combinations (Supplementary Data 3[7]).

Table 3.1. Significant target genes and variants for AMD susceptibility at GWAS loci after eQTL, eCAVIAR and TWAS analyses.

AMD Locus	Lead GWAS SNP	Chromosome: Position	GWAS *P*	eQTL *P*	Target gene(s)	Percentage variability explained	Significant TWAS gene at the locus (FDR)
B3GALTL	rs9564692	13:31821240	3.31×10^{-10}	2.36×10^{-11}*	B3GLCT†	10.47	B3GLCT (1.34×10^{-4})
RDH5/CD63	rs3138141	12:56115778	4.3×10^{-9}	5.69×10^{-19}*	BLOC1S1†	17.8	BLOC1S1 (7.06×10^{-6})
ACAD10	rs61941274	12:112132610	1.07×10^{-9}	8.95×10^{-2}	SH2B3†	0.71	SH2B3 (0.0217)
CFI	rs10033900	4:110659067	5.35×10^{-17}	3.98×10^{-7}*	PLA2G12A	6.17	CFI (3.01×10^{-10}), PLA2G12A (4.30×10^{-10})
PILRB/PILRA	rs7803454	7:99991548	4.76×10^{-9}	3.57×10^{-77}*	PILRB, PILRA, ZCWPW1, TSC22D4	57.51	MEPCE (6.51×10^{-6}), PILRB (2.06×10^{-5})
TMEM97/VTN	rs11080055	17:26649724	1.04×10^{-8}	8.37×10^{-19}*	POLDIP2, SLC13A2**, TMEM199†	17.65	TMEM199 (2.55×10^{-5}), POLDIP2 (8.60×10^{-5})

* eQTL is significant after correction for multiple testing. † Target of causal variant identified by eCAVIAR. ** Retina-specific eQTL.
Only protein-coding genes are shown here. B3GLCT is the new gene symbol for B3GALTL. SH2B3 was identified by GWAS co-localization (eCaviar) and TWAS, two of the three criteria used to identify target genes in our study. Despite its high eQTL p-value, SH2B3 is an excellent biological candidate for AMD because of its association with inflammation [28].

Differential expression analysis of retinal transcriptomes identified 14 genes with and 161 genes without age correction in advanced AMD (FDR$\leq$ 0.20) (Supplementary Data 5[7] and Fig. B.8a). Thus, similarly to results for other complex diseases[88, 89], our differential expression analysis did not detect many gene expression changes, probably because of heterogeneity caused by aging, polygenic inheritance, and environmental factors. We then examined biological pathways by gene set enrichment analysis (GSEA). Immune-regulation and cholesterol-metabolism pathways, previously implicated in GWAS[54], were upregulated in early and advanced AMD, whereas pathways associated with synapse development and function were largely and exclusively downregulated in intermediate AMD (Supplementary Data 5[7]). We note that most of the genes within susceptibility loci for advanced AMD

did not appear to be associated with intermediate AMD despite sufficient power[54]. Thus, intermediate AMD may not be a transitional stage between early and advanced AMD but a separate entity with unique and distinct genetic underpinnings that require further exploration. Furthermore, weighted genecoexpression network analysis of all samples suggested that several of the pathways implicated in AMD operate through closely connected networks in the retina (Fig. B.8b,c and Supplementary Data 6[7]).

3.4 Discussion

GWAS have successfully identified variants at hundreds of loci that contribute to health- and disease-associated traits, thereby defining their broad genetic architecture[90, 91]. Interpretation of GWAS findings, however, remains a major challenge, because a large proportion of associated variants are not in protein-coding genomic regions, and their effects on specific phenotypes often individually appear to be small[92, 93]. eQTL analysis in disease-relevant tissues appears to be a prominent tool for biological interpretation of GWAS loci[85, 94]. Owing to the large sample size, we were able to identify 14,856 eQTLs that modulate retinal gene expression, a substantial proportion of which are not reported in GTEx v7 data. Moreover, we connected the lead GWAS signal to specific target genes at six known AMD-associated loci by at least two lines of evidence. Two of the target genes were validated by three independent methods: B3GLCT encodes a glucosyltransferase[95], and its loss of function leads to Peters plus syndrome[96]; BLOC1S1 encodes a subunit of a multiprotein complex

associated with the biogenesis of an organelle of the endosome-lysosome system[97], and its altered function can affect synaptic function[98]. Thus, altered expression of B3GLCT and BLOC1S1 might affect extracellular-matrix stability or signaling and the degradation of unwanted/recycled proteins, respectively, thereby contributing to AMD pathogenesis. We attribute the lack of obvious target genes at the remaining AMD-GWAS loci to multiple factors, including LD structure, variants affecting expression in trans or in other AMD-relevant tissues (such as retinal pigment epithelium and choroid), and the power of this study. Interpretation of eVariants that regulate multiple genes at a particular locus requires further biological validation.

AMD is notable among complex traits because of its high heritability and large effect sizes for individual GWAS SNPs[54]. We show that variants associated with gene expression across many tissues as eQTLs, as opposed to those with only tissue-specific associations, are enriched in AMD associations despite high tissue specificity of the disease itself (Supplementary Data 3[7] and Fig. B.5g). We hypothesize that, at least in part, such associations reflect larger, more robust effects among the shared eQTLs. Not surprisingly, the retina is the only tissue for which we detected regulation consistently across all six identified SNPs (Supplementary Data 3[7]). In addition, 36 of the 61 retina-identified TWAS candidates were significant (FDR$\leq$0.05) in at least one GTEx tissue. The remaining candidates could not be analyzed because they had no expression or heritability in the GTEx tissues, or they were not replicated in any other tissue. Our results corroborate findings from recent studies[86, 99] and suggest that the best way to increase power for gene discovery through TWAS and similar approaches is to increase the diversity of tissues for greater resolution of the effects

of regulatory variants. We emphasize, however, that eQTL effects detected only in a tissue without biological relevance, but not in a relevant tissue, would be difficult to interpret for disease-specific phenotypes. Although other tissues may show same eQTLs, the retinal effects of eQTLs are more likely to be directly relevant. We suggest that eQTL analyses of retinal pigment epithelium and choroid would further contribute to the understanding of genes involved in AMD pathobiology. AMD associated genes uncovered by TWAS provide additional insights into the relevance of gene regulation to phenotypic consequences in this complex disease.

EyeGEx complements the GTEx project and provides a reference for the biological interpretation of genetic variants associated with common ocular traits, including glaucoma and diabetic retinopathy. Comparative analysis of retinal transcriptomes and eQTLs with the GTEx data should assist in exploring biological questions relating to visual function in syndromic and multifactorial traits.

Chapter 4

Modeling trans-regulatory effect size distribution using summary-level data

4.1 Introduction

Genome-wide association studies (GWAs) have been successful in identifying variants that are implicated in complex traits and diseases. However, most associations that have been found lie in non-protein coding or intergenic regions of the genome and so may be involved in the disease process through gene regulation. Hence, a comprehensive understanding of transcriptome regulation is required to disentangle the biological mechanisms underlying these associations. This has led to the development of approaches based on the integration of both genotype data and gene

expression information to parse out variants involved in gene regulation and how these variants might affect the disease process. These variants, along with the gene or genes whose expression they regulate, are known as expression quantitative loci (eQTLs). These eQTLs can either be cis-acting (affect nearby genes) or trans-acting (affecting genes distal to the variant).

Recently the GTEx consortium[2] released the 8th version of their analyses where they identified both cis and trans-eQTLs across 49 tissues with sample sizes from 85 to 706 individuals. While they were able to identify a large number of cis-eQTLs, the amount of trans-eQTL findings were low across the tissues tested. This low yield of trans associations- driven by their small effect sizes, the small sample size of current studies, and the burden of multiple testing for trans-eQTLs-has been observed in multiple studies[1, 2, 3, 4, 6]. To reduce the burden of multiple testing and so increase statistical power for trans-eQTL analysis, Westra et al[4] used 4,542 trait-associated SNPs in a study consisting of 5,311 individuals to identify 346 unique trans acting variants. More recently, Võsa et al[3] using a similar tactic in a larger study (N = 31,684) and same tissue, focused their trans-eQTL analysis on 10,317 trait-associated SNPs and discovered a third of these variants having a trans-regulatory effect. These two studies by dealing with only trait-associated SNPs and tissues such as Whole Blood-which has a lot of publicly available data- were able to increase their number of findings for trans-eQTLs.

A recent study[100] formally proposed an ominigenic model to account for the dispersion of association signals found for complex traits and diseases across the genome. This model, using the framework of gene-regulatory networks, proposes

that the genetic contributions to complex traits can be partitioned into direct effect from sets of core genes and indirect effects from sets of peripheral genes acting in trans on the set of core genes. Supporting this model are the results found from previous studies[5, 6, 101, 102, 103, 104] which show that a large proportion of heritability (60% - 80%) observed across genes is driven by trans rather than cis effects. These trans regulatory effects have been shown to be more tissue-specific[1, 3] compared to cis effects-which tend to be tissue agnostic-and so are likely implicated in the tissue-specific mechanisms which have an impact on the disease process. However, the number of discoveries from trans-eQTL studies have been small mainly because of their small effect sizes, the small sample sizes of current studies, and the issue of multiple testing[1, 2, 3, 4, 6].

While multiple studies[6, 101, 102, 103, 104] have focused on quantifying the heritability of gene expression that is due to trans-regulatory effects, no study has attempted to estimate how polygenic (i.e., the number of loci that contribute to heritability) trans eQTLs are. In this study, we propose a method that estimates the average polygenicity and quantifies the distribution of trans-eQTL heritability across genes. We do this by modeling the effect size distribution of these trans-regulatory effects across genes. The proposed method extends an earlier approach[105], which is based on a likelihood framework, to deal with multiple outcomes. In our approach, we use summary-level data and circumvent issues related to small sample sizes that are seen in current studies involving trans-eQTLs by marginalizing across genes. In addition, we developed a computationally efficient approach to summarize the data

to deal with the large number of marginal effects typically encountered in trans-eQTL analyses. We show via simulations that the method works, provide a software implementation, and present results based on our application of the method to GTEx V8 data for four tissues. Last, we provide estimates of expected yields in future studies for those tissues.

4.2 Methods

4.2.1 Model

We assume that we have regression coefficients (denoted as $\hat{\beta}^{(M)}$) and their standard errors from the usual marginal model used in eQTL studies. That is, for a given gene, the regression coefficients and their standard errors are obtained from regressing the allele count of one SNP at a time on the expression level of the gene. Hence, assuming we have G genes and a total of K SNPs fitting the trans criteria across all the genes, we end up with $G \times K$ regression coefficients and their respective standard errors. We assume that these estimates are obtained under the scenario where both the outcome and the genotype data have been transformed to have unit variance and mean zero. Let Y_g be the expression level of the gth gene, then the polygenic model for trans eQTLs using this gene can be written $Y_g = \sum_{i=1}^{K} X_k \beta_{gk}^{(J)} + \epsilon$. Y_g and X_k are N $\times$ 1 vectors of gene expression levels and allelic counts for the gth Gene and the kth SNP, respectively, across N subjects. Under this approach, we assume that the distribution of the true effect sizes, based on the joint model, across K_g SNPs

for gene g are i.i.d according to the following distribution

$$\beta_{kg}^{(J)}|\sigma_g^2, \pi_g \sim \pi_g \mathcal{N}(0, \sigma_g^2) + (1 - \pi_g)\delta_0 \tag{4.1}$$

Where both σ_g^2 and π_g vary over genes and δ_0 is the Dirac delta function indicating a fraction, $1 - \pi_g$, of the SNPs have no effect on the expression level of gene g. We assume the following distributions for π_g and σ_g^2

$$\pi_g \sim \text{Beta}(\alpha_0, \beta_0) \text{ and } \sigma_g^2|\pi_g \sim \text{Inverse Gamma}(a_0, b_0) \tag{4.2}$$

set $a_0 = \nu_0/2$ and $b_0 = \nu_0\sigma_0^2/2$. Our goal is to estimate α_0, β_0, ν_0, and σ_0^2.

4.2.2 Composite Likelihood

Given the usual marginal model for trans-eQTLs (I.e, for the kth SNP and gth gene, $Y_g = X_k\beta_{kg}^{(M)} + \epsilon$), and utilizing the following relationship between the marginal effect size, $\beta_{kg}^{(M)}$, and the joint effect size, $\beta_{kg}^{(J)}$, for the kth SNP on the gth gene

$$\beta_{kg}^{(M)} = \sum_{p=1}^{P} \beta_{pg}^{(J)} \rho_{kp} \tag{4.3}$$

where ρ_{kp} is the Pearson correlation coefficient between SNP k and p, it has been shown [105] that under the assumption of independence between LD patterns and probability of a SNP having a non-zero effect on gene, the marginal distribution for $\beta_{kg}^{(M)}$, for SNP k and gene g, can be approximated with the following mixture

distribution

$$\beta_{kg}^{(M)}|\pi_g, \sigma_g^2 \sim \sum_{n_{kg}^{(1)}} f_{N_{kg}^{(1)}|\pi_g,\sigma_g^2}(n_{kg}^{(1)})\mathcal{N}\left(0, \sum_{h=0}^{1} \frac{n_{kg}^{(h)}}{n_{kg}}\sigma_{h,g}^2 \ell_{kg}\right)$$

where $n_{kg}^{(0)} + n_{kg}^{(1)} = n_{kg}$, n_{kg} is the observed number of SNPs in the reference panel that may be "tagged" by the kth SNP with respect to gene g[1], and $n_{kg}^{(1)}$ is the observed number of SNPs, "tagged" by the kth SNP, with independent non-zero effects on gene g based on the joint model. $f_{N_{kg}^{(1)}|\pi_g,\sigma_g^2}(n_{kg}^{(1)})$ follows a binomial distribution with n_{kg} number of trials and π_g probability of success

$$f_{N_{kg}^{(1)}|\pi_g,\sigma_g^2}(n_{kg}^{(1)}) = \frac{n_{kg}!}{n_{kg}^{(1)}! n_{kg}^{(0)}!}\pi_g^{n_{kg}^{(1)}}(1-\pi_g)^{n_{kg}^{(0)}}, \quad n_{kg}^{(1)} = 0, \ldots, n_{kg}$$

Then using the fact that conditional on the true marginal effect size the (marginal) OLS estimate follows a normal distribution, shown below

$$\hat{\beta}_{kg}^{(M)}|\beta_{kg}^{(M)} \sim \mathcal{N}(\beta_{kg}^{(M)}, a + s_{kg}^2)$$

where the factor "a" is introduced to account for possible systematic bias in variance estimates due to effects such as population stratification or cryptic relatedness with respect to gene g, we obtain the likelihood for $\hat{\beta}_{kg}^{(M)}$ by marginalizing (Appendix C)

[1]Recall that trans-eVariants have to be a certain distance away from the gene. Hence, for some SNPs not all variants within its neighborhood will be in $\mathcal{S}_{kg}$. Where $\mathcal{S}_{kg}$ is the set of SNPs in the reference panel that may be "tagged" by the kth SNP with respect to gene g

over the distributions of σ_g^2 and π_g. This gives us

$$\mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kg}^{(M)}) = f^*_{n_{kg}^{(1)}}(0)\mathcal{N}\left(0, a + s_{kg}^2\right) + \sum_{n_{kg}^{(1)} \geq 1} f^*_{n_{kg}^{(1)}}(n_{kg}^{(1)})\mathcal{G}_{\tau_{kg}, s_{kg}^2}(\sigma_0, \nu_0, a) \tag{4.4}$$

$\boldsymbol{\theta} = (\alpha_0, \beta_0, \sigma_0^2, \nu_0, a)$. $f^*_{n_{kg}^{(1)}}(n_{kg}^{(1)})$ is a beta-binomial distribution with parameters α_0, and β_0. $\mathcal{N}\left(0, a + s_{kg}^2\right)$ is the distribution obtained when the kth SNP has no SNP in its neighborhood with a non-zero effect on the gth gene. $\tau_{kg} = \dfrac{n_{kg}^{(1)}}{n_{kg}}\ell_{kg}$, $\ell_{kg} = \sum_{p \in \mathcal{S}_{kg}} \rho_{kp}^2$. I.e., the linkage disequilibrium (LD) score for SNP k with respect to the set of SNPs, $\mathcal{S}_{kg}$, which it tags in the reference panel and satisfy the trans criteria for the gth gene. $\mathcal{G}_{\tau_{kg}, s_{kg}^2}(\sigma_0, \nu_0, a)$ has a complicated form (Appendix C) and is the distribution obtained after marginalizing over a Student's t-distributed mean parameter in a normal distribution. The likelihood in (4.4) is what is obtained after averaging over all the possible states for π_g, and σ_g^2. Here, π_g, and σ_g^2 can be viewed as nuisance parameters which we are not interested in estimating. Ignoring correlation across genes and between SNPs we use the composite likelihood under a working independence assumption. This is given as

$$\mathbf{CL}(\hat{\beta}^{(M)}|\boldsymbol{\theta}) = \prod_{g=1}^{G}\prod_{k=1}^{K_g}\mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kg}^{(M)}) \tag{4.5}$$

Hence the composite log likelihood is

$$\mathbf{cl}(\hat{\beta}^{(M)}) = \sum_{g=1}^{G}\sum_{k=1}^{K_g} \log\left(f^{*}_{n^{(1)}_{kg}}(0)\mathcal{N}\left(0, a + s^2_{kg}\right) + \sum_{n^{(1)}_{kg}\geq 1} f^{*}_{N^{(1)}_{kg}}(n^{(1)}_{kg})\mathcal{G}_{\tau_{kg},s^2_{kg}}\left(\sigma_0, \nu_0, a\right)\right)$$

$$= \sum_{g=1}^{G}\sum_{k=1}^{K_g} \log\left(\mathcal{L}(\boldsymbol{\theta}, \hat{\beta}^{(M)}_{kg})\right)$$

We estimate $\boldsymbol{\theta}$.

Optimization

We can see from (4.5) that the likelihood is maximized over a large set of points. We reduce the computational load by approximating the distribution of the effect sizes per SNP across genes. This is done by binning the effect sizes (Appendix C). This approach is similar to what is done when plotting histograms. We do this for each SNP using a hundred bins, the same bin width across SNPs, and assigning the median effect size to all effect sizes that fall in the same bin. Under this paradigm, and based on 100 bins, (4.5) can be rewritten as

$$\mathbf{CL}(\hat{\beta}^{(M)}|\theta) = \prod_{g=1}^{G}\prod_{k=1}^{K_g} \mathcal{L}(\theta, \hat{\beta}^{(M)}_{kg}) = \prod_{k=1}^{K}\prod_{r=1}^{100}\left[\mathcal{L}(\theta, \hat{\beta}^{(M)}_{kr})\right]^{C_r} \tag{4.6}$$

where C_r is the number of effect sizes that fall within a given bin r.

4.2.3 Estimation

We use the method of differential evolution[106] to obtain the parameter estimates, $\hat{\boldsymbol{\theta}}$, which are global maximizers of the composite likelihood in (4.6) with the constraint that $P(h_g^2 \leq 1) = 1$(Appendix C). This constraint ensures that the parameter estimates, $\hat{\boldsymbol{\theta}}$, are those such that the (unseen) heritability estimates for any gene in our model cannot be greater than 1. In addition, for each SNP we assumed a maximum of 5 neighboring SNPs with a non-zero effect on the outcome.

Estimating $E(h_g^2), Var(h_g^2), E(\pi_g), Var(\pi_g), E(\sigma_g^2)$, **and** $Var(\sigma_g^2)$

$E(\pi_g), Var(\pi_g), E(\sigma_g^2)$, and $Var(\sigma_g^2)$ are obtained using the densities in (C.2). With these we obtain $E(h_g^2)$, and $Var(h_g^2)$ as follows (Appendix C).

$$E(h_g^2) = E(M\pi_g\sigma_g^2) = ME(\sigma_g^2 E(\pi_g|\sigma_g^2))$$

$$= M\mu_\pi\mu_{\sigma_g^2} \tag{4.7}$$

where $\mu_\pi = E(\pi_g)$, $\mu_{\sigma_g^2} = E(\sigma_g^2)$, and M is the total number of SNPs. The variance is

$$Var(h_g^2) = Var(M\pi_g\sigma_g^2) = M^2\left[Var(\sigma_g^2 E(\pi_g|\sigma_g^2)) + E((\sigma_g^2)^2 Var(\pi_g|\sigma_g^2))\right]$$

$$= M^2\left[Var(\sigma_g^2)\mu_\pi^2 + E((\sigma_g^2)^2)Var(\pi_g)\right] \tag{4.8}$$

4.2.4 Variance calculation

Using the same approach seen in [105, 107] the variance for $\hat{\boldsymbol{\theta}}$ is given as

$$\text{var}(\hat{\boldsymbol{\theta}}) = I^{-1}(\boldsymbol{\theta})J(\boldsymbol{\theta})I^{-1}(\boldsymbol{\theta})$$

where

$$I(\boldsymbol{\theta}) = E\left[\sum_{g=1}^{G}\sum_{k=1}^{K_g}\frac{U_{kg}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right], \quad J(\boldsymbol{\theta}) = \text{var}\left\{\sum_{g=1}^{G}\sum_{k=1}^{K_g}U_{kg}(\boldsymbol{\theta})\right\}, \quad U_{kg}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\log\left(\mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kg}^{(M)})\right)$$

We can estimate $I(\boldsymbol{\theta})$ empirically even for correlated data and $J(\boldsymbol{\theta})$ is estimated using the moving block bootstrap approach (Appendix C). Since, we don't have $\boldsymbol{\theta}$ we use the plug in estimate $\hat{\boldsymbol{\theta}}$. For the moving block bootstrap approach, we note that the correlation between the score statistics for any two Genes and SNPs is non-zero if there is both correlation across Genes for a given SNP and across SNPs for a given Gene. Hence, we sum the score statistic across Genes for a given SNP, and apply the moving block approach across SNPs only. This done using overlapping (moving) blocks of size L-defined as the maximum number of SNPs tagged by any SNP in our dataset. We then estimate $\hat{J}(\hat{\boldsymbol{\theta}})$ as the observed variance in the bootstrap samples.

Based on transforms

Recall that the elements of $\hat{\boldsymbol{\theta}}$ are the parameters for the Beta and the Inverse Gamma distributions. Using these parameters we can estimate the uncertainty in the estimates of $E(h_g^2), Var(h_g^2), E(\pi_g), Var(\pi_g), E(\sigma_g^2)$, and $Var(\sigma_g^2)$. This is

done through two applications of the Delta method (Appendix C). In the first step, we obtain the covariance matrix for the joint distribution of $\boldsymbol{\xi} = g(\boldsymbol{\theta}) = \big(E(\pi_g), Var(\pi_g), E(\sigma_g^2), Var(\sigma_g^2)\big)$ shown below

$$\Sigma = g'(\boldsymbol{\theta}) var(\boldsymbol{\theta}) (g'(\boldsymbol{\theta}))^T$$

where $g'(\boldsymbol{\theta})$ is the Jacobian matrix of $g(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ (Appendix C). With this, the covariance matrix for $h(\boldsymbol{\xi}) = (E(h_g^2), Var(h_g^2))$ is given as

$$\Xi = h'(\boldsymbol{\xi}) \Sigma (h'(\boldsymbol{\xi}))^T$$

where $h'(\boldsymbol{\xi})$ is the Jacobian matrix of $h(\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ (Appendix C)

4.2.5 Simulation framework

We generate the summary level results using the following model and simulation scheme. The model used is

$$\hat{\beta}_{kg}^{(M)} = \beta_{kg}^{(M)} + \xi_k + e_{kg}$$

where $\xi_k, k = 1, \ldots, K$ are i.i.d $\mathcal{N}(0, a)$, and the error term $e_g = (e_1, \ldots, e_K)$ follows a multivariate normal distribution with mean zero and covariance matrix $\mathbf{R}$/n. $\mathbf{R}$ is a matrix of LD coefficients for the eQTLs of a given gene, and n being the sample size of the eQTL study with respect to the tissue being used.

The algorith is as follows.

1. With known values for $E(h_g^2)$ and $E(\pi_g)$, specify the variance terms (I.e. $Var(h_g^2), Var(\pi_g)$) using the coefficient of variation, then derive the corresponding values for $E(\sigma_g^2), Var(\sigma_g^2)$ using (4.7, 4.8) (Appendix C).

2. Using these values we obtain $\boldsymbol{\theta}$ using the method of moments, and generate π_g first then h_g^2 (Appendix C). Application of the transformation, $\sigma_g^2 = \dfrac{h_g^2}{M\pi_g}$, to each sample then gives σ_g^2, such that the generated π_g and σ_g^2 follow the specification in (C.2)

3. Generate $\beta_{kg}^{(J)}$ according to the model in (C.1); from this obtain $\beta_{kg}^{(M)}$ using the result in (C.3). ρ_{kp}, the pairwise Pearson correlation coefficient between markers k and p, is estimated using the corresponding sample correlation coefficient in the reference dataset

4. To generate e_g note that $\mathbf{R}$ will be large and that we also need to account for the relationship across genes, so we do the following. We note that the distribution of e_g is the same as the joint distribution of the eQTL summary level statistic under the null of no association between any of the SNPs and the gene. We also note that there is some correlation of the effect sizes expected of given SNP across genes, so we do the following.

   (a) Using the n_{ref} subjects in our 1000 GENOME reference dataset. We generate, independently, standardized pseudo genes $\mathbf{Y}_i = (Y_1, \ldots, Y_G), i = 1, \ldots, n_{ref}$ from a multivariate normal distribution. I.e

$$\mathbf{Y}_i \sim \text{MVN}(0, V)$$

where V can either be estimated from a reference dataset to account for correlation between the G genes or assumed to be an identity matrix to represent independence across the G genes.

(b) Using the genotype data in the reference panel, we calculate the standardized effect sizes for a given gene $\mathbf{u}_g = (u_1, \ldots, u_k)$ for the K SNPs.

(c) Set $e_{kg} = u_{kg}\sqrt{n_{ref}/n}$ to account for the difference in sample size between the reference dataset and the eQTL study. Since $\mathbf{u}_g \sim \mathcal{N}(0, \mathbf{R}/n_{ref})$, then $\mathbf{e}_g \sim \mathcal{N}(0, \mathbf{R}/n)$

We sum $\beta_{kg}^{(M)}, \xi_k, e_{kg}$ and obtain $\hat{\beta}_{kg}^{(M)}$.

Future projection

With the parameters estimated above, we can provide estimates of the future yield for other studies using the observed effect sizes. To this end, let ND_α be the number of Gene-SNP associations obtained at the type 1 error rate of α[2]. Furthermore, assuming that both the expression levels per Gene and allele count per SNP have been transformed to have unit variances and mean zero, then using the per Gene joint effect sizes, $\hat{\beta}_{sg}^{(J)}$ for SNP s and Gene g, we have

$$E(ND_\alpha) \approx G * M \int_{\sigma_g^2 \times \pi_g} \int_{\beta_{sg}^{(J)}} \text{pow}_\alpha(\beta_{sg}^{(J)}) p(\beta_{sg}^{(J)} | \pi_g, \sigma_g^2) p(\pi_g, \sigma_g^2; \hat{\boldsymbol{\theta}}) d\beta_{sg}^{(J)} d\pi_g d\sigma_g^2$$

[2]To account for multiple testing using the Benjamini-Hochberg approach, this is $\alpha/(\#SNPs * \#Genes)$

Where $\text{pow}_\alpha(\beta) = \Phi(-z_{\alpha/2} - \beta\sqrt{n}) + 1 - \Phi(z_{\alpha/2} - \beta\sqrt{n})$, $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution, $z_\alpha = \Phi(1-\alpha)$ is the α^{th} quantile of the standard normal distribution, and $p(\beta_{sg}^{(J)}|\pi_g, \sigma_g^2)p(\pi_g, \sigma_g^2; \hat{\boldsymbol{\theta}})$ is the inferred effect size distribution for a given gene. Using the normal-mixture model with the inverse gamma and beta priors per gene, we have the following result after marginalizing over the gene specific priors

$$E(ND_\alpha) \approx G * M * E(\pi_g) \int_\beta \text{pow}_\alpha(\beta) t(\beta; \sigma_0, \nu_0) d\beta.$$

where $t(\cdot)$ is the generalized student distribution with location parameter zero, scale parameter σ_0, and ν_0 degress of freedom.

In a similar vein we obtain the expected value of the proportion of genetic variance explained by susceptibility SNPs reaching genome-wide significance. After accounting for multiple testing and marginalizing over genes, we have

$$E(GV_\alpha) \approx \int_\beta \beta^2 \text{pow}_\alpha(\beta) f(\beta; \sigma_0, \nu_0) d\beta$$

where

$$f(\beta; \sigma_0, \nu_0) = \frac{\Gamma\left(\frac{\nu_0+3}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)\sqrt{2\pi}} \left(\frac{\nu_0\sigma_0^2}{2}\right)^{\nu_0/2} \left(\frac{2}{x^2+\nu_0\sigma_0^2}\right)^{(\nu_0+3)/2}$$

4.2.6 GTEx V8 data

Data generation and processing for the summary level data for the four GTEx tissues (Skeletal Muscle, Adipose, Testis, and Whole Blood) were described extensively in

Aguet et al., 2019[2]. Briefly, each SNP and Gene were defined to be in trans distance of each other if they were on different chromosomes. The covariates in model were the first five genotype PCs, WGS sequencing platform (HiSeq 2000 or HiSeq X), WGS library construction protocol (PCR-based or PCR-free), donor sex, and PEER factors optimized for tissue sample size. Additional quality control for trans-eQTL mapping involved the removal of any variant with mappability < 1, variants in the MHC region, variants which are not HapMap 3 SNPs, the exclusion of any gene with mappability < 0.8, and the removal of any variant-gene pair where the gene cross-maps with any other gene within 1Mb distance of the variant. Hence we ended up with summary level data for approximately 1.11 million SNPs across all tissues and between approximately 17,000 to 24,000 Genes across the four tissues. The effect sizes for each SNP across Genes were summarized (nbins = 100) using the approach defined in (4.6).

4.2.7 Estimation of LD-score (ℓ_{kg}) and number of tagged SNPs (n_{kg})

The LD score (defined here as the sum of the squared Pearson correlation coefficients above a given threshold, $r^2 \geq \rho$, between a given SNP and all SNPs within a window of ws) and the number of SNPs tagged (here defined as the number SNPs within the window and having $r^2 \geq \rho$ with the SNP of interest) by each SNP used in our simulations were obtained using 489 European individuals from the 1000 Genomes reference panel as described previously[105]. For the GTEx V8 data, ℓ_{kg} and n_{kg} were estimated using the GCTA software[76]. In both sets (i.e., simulation datasets

and GTEx V8 data) of estimates ws = 1Mbps, and $\rho = 0.1$.

4.3 Results

4.3.1 Simulation studies

Our model consists of five parameters of interest (see Methods). Different combinations of these parameters along with the number of SNPs provide estimates of six different quantities. These quantities for heritability, polygenicity, and per-SNP heritability are the averages and the dispersion of each across genes. Implicit in our estimates is the tissue being used. Hence, for each tissue we may obtain different results. Of these six quantities, the average heritability ($\mathrm{E}(h_g^2)$), the dispersion of heritability across genes ($\mathrm{SD}(h_g^2)$), and the average polygenicity ($\mathrm{E}(\pi_g)$) are of main interest since results from our simulations about the remaining parameters were unclear. Hence, we treat the remaining three quantities as nuisance parameters which are needed to provide accurate estimates of $\mathrm{E}(h_g^2)$, $\mathrm{SD}(h_g^2)$, and $\mathrm{E}(\pi_g)$ but are not of direct interest to us.

We ran simulations using summary-level statistics while accounting for the LD patterns observed across SNPs. The summary-level statistics were generated assuming 20,000 genes and using HapMap 3 SNPs with minor allele frequencies (MAF) of at least 5%. Due to the long computation time of our approach we generated data using SNPs from either chromosome 22 only (#SNPS = 15,584) or from both chromosome 21 and 22 (#SNPS = 31,195). The result in each scenario in our simulations is averaged over 50 datasets, and estimation is carried out using 100 bins for each

dataset. In addition, we assumed that each SNP has maximum of 5 neighbouring SNPs, on average across genes, with a non-zero effect on the outcome (see Methods).

We first assessed the accuracy of our method under different amounts of average polygenicity ($E(\pi_g) = 2.5\%$, and $E(\pi_g) = 30\%$) assuming a small per SNP heritability on average across genes. The average (across genes) per SNP heritability, $\mathrm{E}(\sigma_g^2)$, was calculated by first setting $\mathrm{E}(h_g^2)$ to 0.12, $\mathrm{E}(\pi_g)$ to 30% and assuming we used a full set of genome-wide SNPs - approximately 1 Million HapMap 3 SNPs at MAF 5% (see Methods). This produced an average per SNP heritability across genes as 4e-7 (i.e., $\mathrm{E}(\sigma_g^2) = 4\text{e-}7$) which was then fixed and used with different amounts of average polygenicity ($\mathrm{E}(\pi_g)$: 2.5%, and 30%) . Hence, the total heritability on average across genes using SNPs in chromosome 22 only was small ($\mathrm{E}(h_g^2)$= 1.6e-4, $\mathrm{E}(\pi_g) = 2.5\%$; and $\mathrm{E}(h_g^2) = 1.9\text{e-}3$, $\mathrm{E}(\pi_g) = 30\%$). With these values, we then specified the dispersion terms, $\mathrm{SD}(h_g^2)$ and $\mathrm{SD}(\pi_g)$, using the coefficient of variation. This was set to 30%, which reflects a belief of moderate dispersion in relation to the mean (i.e., $\mathrm{E}(h_g^2)$, and $\mathrm{E}(\pi_g)$). Since $\mathrm{E}(\sigma_g^2)$, and by extension $\mathrm{E}(h_g^2)$, was small, this would mean that we are using SNPs, in chromosome 22 to estimate genome-wide effects. Hence, we inflated the sample size by a factor of 64 (this roughly equates to the number of common(MAF$\geq$ 5%) SNPs in HapMap3 divided by the number of common HapMap3 SNPs in chromosome 22 only) to ensure that we have adequate power for this set of simulations. Doing this reduced the impact of the residual error (see Methods). We found that the bias for $\mathrm{E}(h_g^2)$, $\mathrm{SD}(h_g^2)$, and $\mathrm{E}(\pi_g)$ reduced as the sample size increased (Figure 4.1).

Next, we assumed that the set of SNPs being used-chromosome 22 SNPs-explained

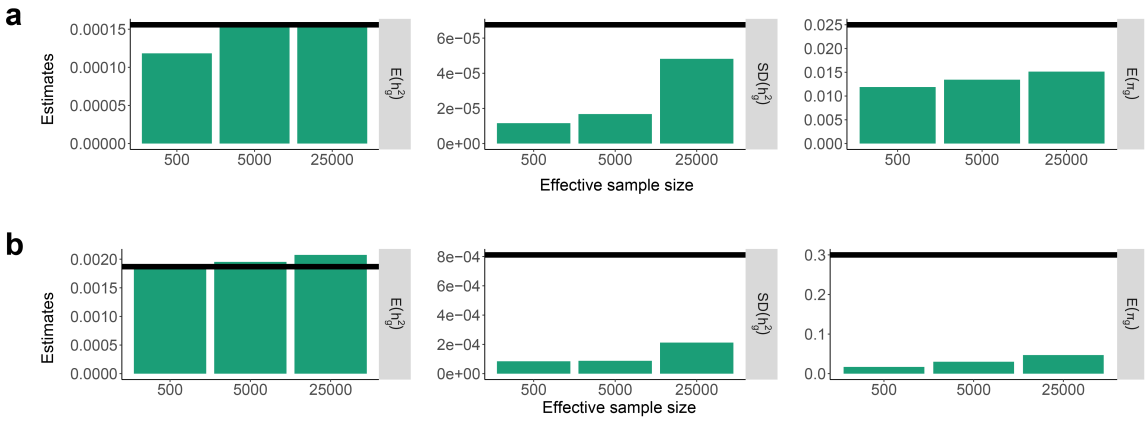


Fig. 4.1. Comparison of estimates obtained using averaged across 50 datasets at a low per SNP heritability (4e-7); We show results based on an average polygenicity of 2.5% (**a**) and an average polygenicity of 30% (**b**). Effective sample size is ($N \times$ #SNPs in Chr 22)/(#SNPs genomewide). Horizontal black lines correspond to the truth. Note that the y-axis in each subplot are in different scales.

the same amount of variability in the outcome as what we would see if we had used a full set of genome-wide SNPs instead. Hence, for a given amount of polygenicity, this corresponds to SNPs in our datasets having a larger effect on the outcome than what we would realistically see genome-wide. Our results generally remained the same for $\mathrm{E}(h_g^2)$, $\mathrm{SD}(h_g^2)$, and $\mathrm{E}(\pi_g)$ when we increased the per-SNP heritability and did not inflate the sample sizes used in our simulations (Figure C.1). Comparing the main quantities of interest ($\mathrm{E}(h_g^2)$, $\mathrm{SD}(h_g^2)$, and $\mathrm{E}(\pi_g)$) as the average per SNP heritability increases (5e-5 vs 4e-7), we found that as the average per SNP heritability increases we obtain better results even after inflating the sample size when the average per SNP heritability is small (Figure C.2).

We then investigated the effect of increasing the number of SNP used for estimation on the quantities of interest for fixed values of $\mathrm{E}(h_g^2)$, $\mathrm{SD}(h_g^2)$ and $(\mathrm{E}(\pi_g))$. Hence, with these fixed quantities the average per SNP-heritability, $\mathrm{E}(\sigma_g^2)$, would be

smaller when we have more SNPs. The comparisons were done using SNPs from chromosome 22 only and all SNPs from chromosome 21 and 22. We found that with a larger SNP set, we obtained estimates of SD(h_g^2) that are less biased (Figure C.3).

To see how our method works as the variability of π_g/σ_g^2 increases/decreases across genes, we fixed E(h_g^2), E(π_g), and SD(h_g^2) while we varied the coefficient of variation for π_g from 20% to 80% and ran simulations using chromosome 22 SNPS only. Note that since E(h_g^2), E(π_g), and SD(h_g^2) are fixed, our results are also equivalent to decreasing the coefficient of variation for σ_g^2 from 80% to 20% (see Methods). We saw negligible differences in the mean terms, E(h_g^2) and E(π_g), as the variability of π_g increased across genes (or as the variability of σ_g^2 decreased across genes). However, the variance of h_g^2 was better estimated when π_g varied less across genes (or as σ_g^2 varied more across genes. Figure C.4). Similarly, when we ran simulations allowing the variability π_g and σ_g^2 to increase together, we found negligible differences (Figure C.5) in the estimates for E(h_g^2) and E(π_g) with smaller values of SD(h_g^2) providing slightly better estimates.

4.3.2 Application to four tissues in GTEx V8

We applied our model, using summary-level results from GTEx V8 data, to four tissues (Adipose, Testis, Skeletal Muscle and Whole blood) and estimated the trans-polygenicity and trans-heritability of gene expression. For each tissue, we obtained summary level results based on trans-eQTL analysis for each tissue as described previously[2](see Methods). In total we ended up with approximately 1.1 million SNPs (MAF $\geq$ 5%) and between approximately 17,000 to 24,000 genes across the

a

Tissue	h^2_g (Mean ± SD)	E (π_g)	N
Adipose Subcutaneous	0.32 ± 0.19	0.35	581
Muscle Skeletal	0.20 ± 0.17	0.28	706
Whole Blood	0.24 ± 0.18	0.33	670
Testis	0.77 ± 0.05	0.23	322

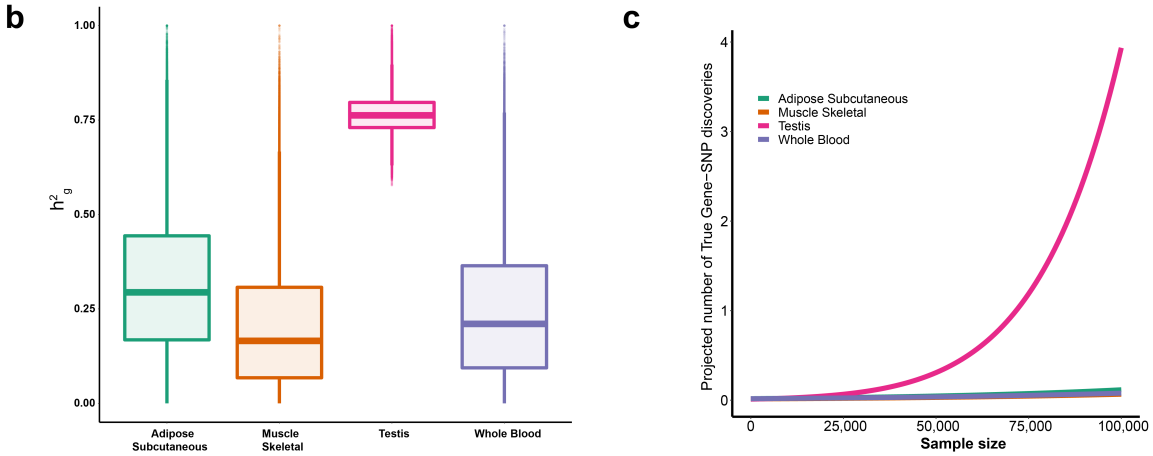


Fig. 4.2. Results from four GTEx V8 Tissues. (**a**), A summary of trans-heritability and trans-polygenicity estimates across genes for each tissue. (**b**), Boxplots showing the distribution of trans-heritability across genes for each tissue. The boxplots depict the median, and the lower and upper hinges correspond to the first and third quartiles, respectively. Outlying data are represented by individual points that extend beyond 1.5 x interquartile range below the first quartile or above the third quartile. (**c**), Projected yield of future studies colored by tissue. These results are based on power calculations for discovery after Bonferroni correction ($P = 2.24 \times 10^{-12}$) using 20,000 genes and 1.1 million SNPs for all tissues.

four tissues. The sample sizes per tissue ranged from 322 for Testis to 706 for Muscle.

Across tissues, the estimates for heritability were the highest for Testis and lowest for Skeletal Muscle. Furthermore, the estimates for the dispersion of heritability across genes was similar for all tissues except Testis (Figure 4.2a,b). Estimates of the average polygenicity across tissues (Figure 4.2a) were much higher than what have been previously reported for complex traits[105, 108]. Projected yields, using observed effect sizes, of future studies showed patterns consistent with trans-heribility estimates(Figure 4.2c). In particular, the rate at which the projected number of discoveries increased as sample size increased was highest for Testis. However, across all tissues in this study the yields were minimal even at high sample sizes.

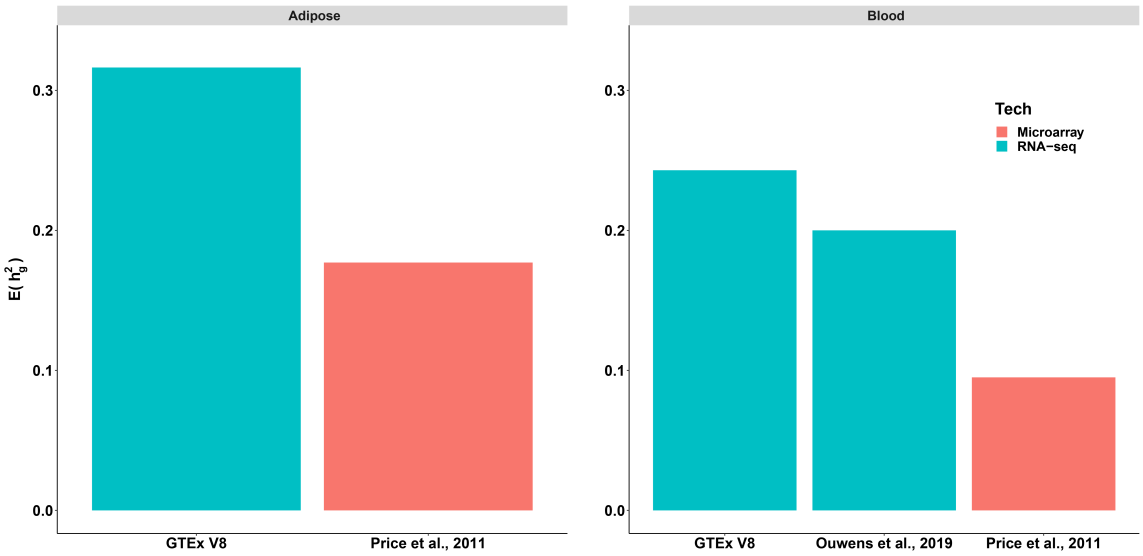


Fig. 4.3. Mean heritability of trans-effects across studies. We show barcharts comparing our estimates to what has been observed. The plot on the left is based on the Adipose tissue, while the one on the right is based on Whole blood. The colors represent the different technology used by each study to generate gene expression data.

Comparing our estimates to what is observed in literature, we found that for

tissues where comparisons exist (Adipose and Whole Blood), our estimates for the mean heritability were consistently larger (Figure 4.3), regardless of the technology used to generate the gene expression data and the sample size of the study. Furthermore, consistent with has been reported previously[102], comparisons across different technologies used in each study to generate the gene expression data showed that estimates from studies which used rna-seq were consistently higher despite having a smaller sample size.

4.4 Discussion

In this study we model the effect size distribution of trans effects using a likelihood framework derived from a two component mixture model. We address the issue of small sample sizes seen in current studies by marginalizing over genes. We provide estimates and present summary statistics concerning the distribution of h_g^2, that are due to trans regulatory effects, across genes. We show that the estimates can be quite different across tissues and that our estimates correspond with what have been reported previously(Figure 4.3)[2, 6, 102]. In addition, we provide estimates of the average polygenicity across genes for each tissue. Comparison of these estimates show that relative to complex traits and diseases, trans-regulatory effects are much more polygenic[105, 108]. However, given that the estimates of the average trans-eQTL heritability across genes are not too different from that seen for complex traits and diseases, this implies that the effect sizes for trans-regulatory effects are much smaller. Coupled with the fact that accounting for multiple testing is a necessity

in trans-eQTL studies and the relatively small sample sizes of current studies, we see that at least when compared to complex traits and disease, identification of trans-eQTLs will be a substantially more difficult task. This conclusion was further supported when our estimates of yields in future studies were minimal even at large sample sizes.

Our simulation studies, which were run using small SNP sets because of the large computational load of our method, displayed patterns that are consistent with what was seen in a closely related approach[105]. In detail, we obtained estimates of mean heritability which showed little to no bias as the sample size increased. In addition, increases in the number of SNPs used for estimation had a negligible impact on the mean terms while decreasing the bais for the dispersion term (Figure C.3). The remaining parameters, i.e., the dispersion of heritability and the mean polygenicity, were consistently underestimated across the various simulation scenarios that we used. Some of these results were due to the smaller effect sizes used which correspond to more realistic scenarios but used only a small subset of SNPs for estimation. Nevertheless, we observed an improvement in our estimates for given amounts of polygenicity when the average heritability increased (Figure C.2). Additional possible reasons for the underestimation involved 1) assumptions we made about the number of neighbouring SNPs that have a non-zero effect on the outcome on average across genes (we assumed that each SNP has a maximum of 5), and 2) additional biases due to SNPs with small effects being indistinguishable from zero. In simulations with small SNP sets, we observed that allowing increases in the number of neighbouring SNPs that have a non-zero effect on the outcome on average across

genes improved our estimates of polygenicity. However, the bias did not go to zero and it's unclear what effect such increases will have when dealing with the full set of genome-wide SNPs. We hypothesize that some of the current issues are based on the fact that our model estimates a lot of parameters, some of which may not be well-identified in our simulations using a limited SNP size.

Comparisons between our results and reported estimates are complicated because of the different technologies used for gene quantification, the different quality control measures taken, differences in tissue collection, sampling variation, and other methodological differences across studies[5, 109]. In spite of this, our estimate for Whole Blood was similar to a previous result[102] using rna-seq data in the same tissue. Moreover the pattern of heritability between tissues was consistent with previous results[103] with the average heritability in the Adipose tissue being larger than the corresponding estimate in the Whole Blood tissue. Concerning the impact of technological differences used for gene quantification, Ouwens et al., 2019 showed that heritability estimates obtained using rna-seq rather than microarray data, in the same population and tissue, were larger likely because rna-seq is better able to capture variation in low to moderately expressed genes[102, 110]. While our result for Whole Blood was larger than the estimate obtained by Ouwens et al., 2019, we note that the sequencing coverage in GTEx V8 data was much larger (minimum coverage; 50M vs 15M paired end reads). Hence, the resolution of gene-expression in GTEx V8 rna-seq data is better and likely provides better information about variation in low to moderately expressed.

Limitations of this study include the choice of r^2 threshold and window size used

to generate both the LDScores and to specify the number of SNPs tagged for each SNP. We didn't investigate this since a previous study[105] using a closely related method already evaluated the impact of this on the estimated results. Instead, we used an r^2 threshold of 0.1 and a 1 MB window size. A combination which has been shown to produce estimates of effect-size distributions which are comparable to those obtained from alternative combinations that are optimal under different scenarios. For the simulation study, we were limited by the small number of SNPs used for evaluation. As a result, we weren't able to evaluate the accuracy of the estimates of the standard error for each parameter of interest. While our simulation results and results from a closely related method in another study[105] show that the parameters (average and dispersion) relating to polygenicity are consistently underestimated, we argue that the estimates especially that of the average polygenicity still provide a limited use, and can serve as the lower bound on the average polygenicity of trans-effects across genes.

Another possible limitation of the proposed method includes the assumption that effect sizes are independent of allele frequencies and local LD patterns of SNPs. While it has been shown that this may lead to underestimates of heritability for complex traits and diseases[111], it's unclear if this problem will persist when dealing with gene expression data. In addition, we note that due to the complexity of our model the average runtime, using 40 CPUs with a total of 30 GB RAM, for one million SNPs using 100 bins and assuming a maximum of 5 neighbouring SNPs with a non-zero effect on average on the outcome was between one to two weeks depending on the tissue of interest.

In summary, we have developed a method to model the effect size distribution of trans regulatory effects. We show in simulations that the method works and provide estimates for four tissues in GTEx V8. Furthermore, we provide a way to estimate yields in future studies which can then be utilized during the design and planning phase. While we did not evaluate the accuracy of our method in such scenarios, we note that the method can easily be adapted to work with summary level data from any analysis involving multiple phenotypes such as proteomics, metabolomics, microbiome studies, epigenomics, and other phenome-related approaches. Despite the limitations in the study, our study advances our understanding of the genetic architecture of trans-regulatory effects.

Chapter 5

Conclusions

In this final chapter, we would like to summarize the contributions of this body of work, and its possible impacts in public health.

5.1 Addressing the bias observed in estimates of absolute cell fractions using RNA-Seq data (Chapter 2)

Using several different reference datasets including a bulk/homogenate dataset with paired DNAm and RNA-seq data from the nucleus accumbens (NAc) from 200+ deceased individuals, we show that RNA-based deconvolution for just two cell populations - neurons and non-neurons - largely fails to estimate the underlying cellular composition of bulk human brain tissue across a variety of algorithms and strategies. We quantified the diverse range of neuronal fractions estimated by several popular

algorithms to better understand the effects of reference cell type-specific expression profiles and differences in cell size and/or activity profiles on deconvolution. We specifically examined the common scenario of performing RNA deconvolution using cell type-specific reference datasets that can be fundamentally different from user-provided homogenate tissue target datasets, for example differing in profiled brain region, sequencing technology and/or cellular compartment. These problems are likely magnified in human brain tissue compared to suspended cells like blood, where deconvolution strategies are more easily validated against true cell fractions obtained by routine complete cell counts[11]. We lastly emphasize caution when performing RNA-based deconvolution using many cell types (i.e., more finely-partitioned cell classes) without having the ability to validate cell counts on at least a subset of samples. We therefore provide several recommendations for performing RNA-based deconvolution in bulk human brain gene expression data.

5.2 Integrative analysis using both gene expression and genetic data to provide valuable insights in the disease progression for Age-related Macular Degeneration(AMD) (Chapter 3)

Utilizing data from cis regulatory effects in the retina and information about trait-associated SNPs from AMD GWAs, we were able to connect the lead GWAS signal to specific target genes at six known AMD loci by multiple lines of evidence. Both

of the target genes, B3GLCT and BLOC1S1, that were identified by three independent methods in our study have been replicated in a related tissue, retinal pigment epithelium (RPE), as being target genes for AMD[112] . Furthermore, functional validation experiments in zebrafish, where expression levels for BLOC1S1 were over expressed led to mild impaired ocular phenotypes[113]. We also showed that even for a multifactorial disease like AMD with high tissue specificity, variants associated with gene expression across many tissues as eQTLs, as opposed to those with tissue-specific associations only, are more likely to show disease association.

5.3 Modeling trans-regulatory effect size distribution (Chapter 4)

We developed a likelihood framework to estimate the average polygenicity and quantify the distribution of trans-eQTL heritability across genes. We present a software application and apply the method to summary level data obtained for four tissues from the Genotype-Tissue Expression(GTEx) Consortium. In addition, we provide lower bound estimates of the average polygenicity of trans-regulatory effects in these four tissues, and use results from our model fit to make projections about future studies. Results from our model are consistent with the current sparsity of trans-eQTL results.

References

[1] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, October 2017. ISSN 1476-4687. doi: 10.1038/nature24277. URL https://www.nature.com/articles/nature24277. Number: 7675 Publisher: Nature Publishing Group.

[2] Franois Aguet, Alvaro N Barbeira, Rodrigo Bonazzola, Andrew Brown, Stephane E Castel, Brian Jo, Silva Kasela, Sarah Kim-Hellmuth, Yanyu Liang, Meritxell Oliva, Princy E Parsana, Elise Flynn, Laure Fresard, Eric R Gaamzon, Andrew R Hamel, Yuan He, Farhad Hormozdiari, Pejman Mohammadi, Manuel Muoz-Aguirre, YoSon Park, Ashis Saha, Ayellet V Segr, Benjamin J Strober, Xiaoquan Wen, Valentin Wucher, Sayantan Das, Diego Garrido-Martn, Nicole R Gay, Robert E Handsaker, Paul J Hoffman, Seva Kashin, Alan Kwong, Xiao Li, Daniel MacArthur, John M Rouhana, Matthew Stephens, Ellen Todres, Ana Viuela, Gao Wang, Yuxin Zou, The GTEx Consortium, Christopher D Brown, Nancy Cox, Emmanouil Dermitzakis, Barbara E Engelhardt, Gad Getz, Roderic Guigo, Stephen B Montgomery, Barbara E Stranger, Hae Kyung Im, Alexis Battle, Kristin G Ardlie, and Tuuli Lappalainen. The

GTEx Consortium atlas of genetic regulatory effects across human tissues. preprint, Genetics, October 2019. URL http://biorxiv.org/lookup/doi/10.1101/787903.

[3] Urmo Vsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, Natalia Pervjakova, Isabel Alvaes, Marie-Julie Fave, Mawusse Agbessi, Mark Christiansen, Rick Jansen, Ilkka Seppl, Lin Tong, Alexander Teumer, Katharina Schramm, Gibran Hemani, Joost Verlouw, Hanieh Yaghootkar, Reyhan Snmez, Andrew A. Andrew, Viktorija Kukushkina, Anette Kalnapenkis, Sina Reger, Eleonora Porcu, Jaanika Kronberg-Guzman, Johannes Kettunen, Joseph Powell, Bernett Lee, Futao Zhang, Wibowo Arindrarto, Frank Beutner, BIOS Consortium, Harm Brugge, i2QTL Consortium, Julia Dmitrieva, Mahmoud Elansary, Benjamin P. Fairfax, Michel Georges, Bastiaan T Heijmans, Mika Khnen, Yungil Kim, Julian C. Knight, Peter Kovacs, Knut Krohn, Shuang Li, Markus Loeffler, Urko M Marigorta, Hailang Mei, Yukihide Momozawa, Martina Mller-Nurasyid, Matthias Nauck, Michel Nivard, Brenda Penninx, Jonathan Pritchard, Olli Raitakari, Olaf Rotzschke, Eline P. Slagboom, Coen D.A. Stehouwer, Michael Stumvoll, Patrick Sullivan, Peter A.C. 't Hoen, Joachim Thiery, Anke Tnjes, Jenny van Dongen, Maarten van Iterson, Jan Veldink, Uwe Vlker, Cisca Wijmenga, Morris Swertz, Anand Andiappan, Grant W. Montgomery, Samuli Ripatti, Markus Perola, Zoltan Kutalik, Emmanouil Dermitzakis, Sven Bergmann, Timothy Frayling, Joyce van Meurs, Holger Prokisch, Habibul Ahsan, Brandon Pierce,

Terho Lehtimki, Dorret Boomsma, Bruce M. Psaty, Sina A. Gharib, Philip Awadalla, Lili Milani, Willem H. Ouwehand, Kate Downes, Oliver Stegle, Alexis Battle, Jian Yang, Peter M. Visscher, Markus Scholz, Gregory Gibson, Tnu Esko, and Lude Franke. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. preprint, Genomics, October 2018. URL http://biorxiv.org/lookup/doi/10.1101/447367.

[4] Harm-Jan Westra, Marjolein J Peters, Tnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, Alexandra Zhernakova, Daria V Zhernakova, Jan H Veldink, Leonard H Van den Berg, Juha Karjalainen, Sebo Withoff, Andr G Uitterlinden, Albert Hofman, Fernando Rivadeneira, and Peter A C 't Hoen. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, October 2013. ISSN 10614036. doi: 10.1038/ng.2756. URL http://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=90429631&site=ehost-live&scope=site.

[5] Fred A Wright, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, Abdel Abdellaoui, Sandra Batista, Casey Butler, Guanhua Chen, Ting-Huei Chen, David D'Ambrosio, Paul Gallins, Min Jin Ha, Jouke Jan Hottenga, and Shunping Huang. Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*, 46(5):430–437, May 2014. ISSN 10614036.

doi: 10.1038/ng.2951. URL http://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=95774119&site=ehost-live&scope=site.

[6] Elin Grundberg, Kerrin S. Small, sa K. Hedman, Alexandra C. Nica, Alfonso Buil, Sarah Keildson, Jordana T. Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, James Nisbett, Magdalena Sekowska, Alicja Wilk, So-Youn Shin, Daniel Glass, Mary Travers, Josine L. Min, Sue Ring, Karen Ho, Gudmar Thorleifsson, Augustine Kong, Unnur Thorsteindottir, Chrysanthi Ainali, Antigone S. Dimas, Neelam Hassanali, Catherine Ingle, David Knowles, Maria Krestyaninova, Christopher E. Lowe, Paola Di Meglio, Stephen B. Montgomery, Leopold Parts, Simon Potter, Gabriela Surdulescu, Loukia Tsaprouni, Sophia Tsoka, Veronique Bataille, Richard Durbin, Frank O. Nestle, Stephen O'Rahilly, Nicole Soranzo, Cecilia M. Lindgren, Krina T. Zondervan, Kourosh R. Ahmadi, Eric E. Schadt, Kari Stefansson, George Davey Smith, Mark I. McCarthy, Panos Deloukas, Emmanouil T. Dermitzakis, and Tim D. Spector. Mapping cis - and trans -regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089, October 2012. ISSN 1546-1718. doi: 10.1038/ng.2394. URL https://www.nature.com/articles/ng.2394.

[7] Rinki Ratnapriya, Olukayode A. Sosina, Margaret R. Starostik, Madeline Kwicklis, Rebecca J. Kapphahn, Lars G. Fritsche, Ashley Walton, Marios Arvanitis, Linn Gieser, Alexandra Pietraszkiewicz, Sandra R. Montezuma, Emily Y. Chew, Alexis Battle, Gonalo R. Abecasis, Deborah A.

Ferrington, Nilanjan Chatterjee, and Anand Swaroop. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nature Genetics*, 51(4):606, April 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0351-9. URL https://www.nature.com/articles/s41588-019-0351-9.

[8] Farhad Hormozdiari, Martijn vandeBunt, AyelletV. Segr, Xiao Li, JongWhaJ. Joo, Michael Bilow, JaeHoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics*, 99(6):1245–1260, December 2016. ISSN 0002-9297. doi: 10.1016/j.ajhg.2016.10.003. URL http://www.sciencedirect.com/science/article/pii/S0002929716304396.

[9] Spyros Darmanis, Steven A. Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M. Shuer, Melanie G. Hayden Gephart, Ben A. Barres, and Stephen R. Quake. A survey of human brain transcriptome diversity at the single cell level. *PNAS*, 112(23):7285–7290, June 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1507125112. URL https://www.pnas.org/content/112/23/7285. Publisher: National Academy of Sciences Section: Biological Sciences.

[10] Alexander R. Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F. Clark. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLOS ONE*, 4(7):e6098, July 2009. ISSN 1932-6203. doi: 10.1371/journal.

pone.0006098. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006098. Publisher: Public Library of Science.

[11] Aaron M. Newman, Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn, and Ash A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, May 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3337. URL https://www.nature.com/articles/nmeth.3337. Number: 5 Publisher: Nature Publishing Group.

[12] Julien Bryois, Alfonso Buil, Pedro G. Ferreira, Nikolaos I. Panousis, Andrew A. Brown, Ana Viuela, Alexandra Planchon, Deborah Bielser, Kerrin Small, Tim Spector, and Emmanouil T. Dermitzakis. Time-dependent genetic effects on gene expression implicate aging processes. *Genome Res*, 27(4):545–552, April 2017. ISSN 1088-9051. doi: 10.1101/gr.207688.116. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5378173/.

[13] Kerrin S Small, sa K Hedman, Elin Grundberg, Alexandra C Nica, Gudmar Thorleifsson, Augustine Kong, Unnur Thorsteindottir, So-Youn Shin, Hannah B Richards, Nicole Soranzo, Kourosh R Ahmadi, Cecilia M Lindgren, Kari Stefansson, Emmanouil T Dermitzakis, Panos Deloukas, Timothy D Spector, Mark I McCarthy, the MuTHER Consortium, the GIANT Consortium, the MAGIC Investigators, and the DIAGRAM Consortium. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature Genetics*, 43(6):561–564, June 2011. ISSN 1546-1718. doi:

10.1038/ng.833. URL https://www.nature.com/articles/ng.833. Number: 6 Publisher: Nature Publishing Group.

[14] Andrew E. Jaffe and Rafael A. Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31, February 2014. ISSN 1474-760X. doi: 10.1186/gb-2014-15-2-r31. URL https://doi.org/10.1186/gb-2014-15-2-r31.

[15] Eugene Andres Houseman, William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, and Karl T. Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, May 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-86. URL https://doi.org/10.1186/1471-2105-13-86.

[16] Carolina M. Montao, Rafael A. Irizarry, Walter E. Kaufmann, Konrad Talbot, Raquel E. Gur, Andrew P. Feinberg, and Margaret A. Taub. Measuring cell-type specific differential methylation in human brain tissue. *Genome Biology*, 14(8):R94, August 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-8-r94. URL https://doi.org/10.1186/gb-2013-14-8-r94.

[17] Karin B. Michels, Alexandra M. Binder, Sarah Dedeurwaerder, Charles B. Epstein, John M. Greally, Ivo Gut, E. Andres Houseman, Benedetta Izzi, Karl T. Kelsey, Alexander Meissner, Aleksandar Milosavljevic, Kimberly D. Siegmund, Christoph Bock, and Rafael A. Irizarry. Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*,

10(10):949–955, October 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2632. URL https://www.nature.com/articles/nmeth.2632. Number: 10 Publisher: Nature Publishing Group.

[18] Jerry Guintivano, Martin J. Aryee, and Zachary A. Kaminsky. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, 8 (3):290–302, March 2013. ISSN 1559-2294. doi: 10.4161/epi.23924. URL https://doi.org/10.4161/epi.23924.

[19] Eugene Andres Houseman, John Molitor, and Carmen J. Marsit. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, May 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu029. URL https://academic.oup.com/bioinformatics/article/30/10/1431/266465. Publisher: Oxford Academic.

[20] Francisco Avila Cobos, Jo Vandesompele, Pieter Mestdagh, and Katleen De Preter. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11):1969–1979, June 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty019. URL https://academic.oup.com/bioinformatics/article/34/11/1969/4813737. Publisher: Oxford Academic.

[21] Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G. Burchard, Eleazar Eskin, James Zou, and Eran Halperin. Sparse PCA corrects for cell type heterogeneity in epigenome-wide

association studies. *Nature Methods*, 13(5):443–445, May 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3809. URL https://www.nature.com/articles/nmeth.3809. Number: 5 Publisher: Nature Publishing Group.

[22] Amanda J. Price, Leonardo Collado-Torres, Nikolay A. Ivanov, Wei Xia, Emily E. Burke, Joo Heon Shin, Ran Tao, Liang Ma, Yankai Jia, Thomas M. Hyde, Joel E. Kleinman, Daniel R. Weinberger, and Andrew E. Jaffe. Divergent neuronal DNA methylation patterns across human cortical development reveal critical periods and a unique role of CpH methylation. *Genome Biology*, 20(1):196, September 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1805-1. URL https://doi.org/10.1186/s13059-019-1805-1.

[23] Neeltje E. M. van Haren, Hugo G. Schnack, Wiepke Cahn, Martijn P. van den Heuvel, Claude Lepage, Louis Collins, Alan C. Evans, Hilleke E. Hulshoff Pol, and Ren S. Kahn. Changes in cortical thickness during the course of illness in schizophrenia. *Arch. Gen. Psychiatry*, 68(9):871–880, September 2011. ISSN 1538-3636. doi: 10.1001/archgenpsychiatry.2011.88.

[24] Michael D. Nelson, Andrew J. Saykin, Laura A. Flashman, and Henry J. Riordan. Hippocampal Volume Reduction in Schizophrenia as Assessed by Magnetic Resonance Imaging: A Meta-analytic Study. *Arch Gen Psychiatry*, 55 (5):433–440, May 1998. ISSN 0003-990X. doi: 10.1001/archpsyc.55.5.433. URL https://jamanetwork.com/journals/jamapsychiatry/fullarticle/203854. Publisher: American Medical Association.

[25] Alexey Kozlenkov, Junhao Li, Pasha Apontes, Yasmin L. Hurd, William M.

Byne, Eugene V. Koonin, Michael Wegner, Eran A. Mukamel, and Stella Dracheva. A unique role for DNA (hydroxy)methylation in epigenetic regulation of human inhibitory neurons. *Science Advances*, 4(9):eaau6190, September 2018. ISSN 2375-2548. doi: 10.1126/sciadv.aau6190. URL `https://advances.sciencemag.org/content/4/9/eaau6190`. Publisher: American Association for the Advancement of Science Section: Research Article.

[26] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R. Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10(1):1–9, January 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-08023-x. URL `https://www.nature.com/articles/s41467-018-08023-x`. Number: 1 Publisher: Nature Publishing Group.

[27] Shahin Mohammadi, Neta Zuckerman, Andrea Goldsmith, and Ananth Grama. A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. *Proceedings of the IEEE*, 105(2):340–366, February 2017. ISSN 1558-2256. doi: 10.1109/JPROC.2016.2607121. Conference Name: Proceedings of the IEEE.

[28] Ting Gong and Joseph D. Szustakowski. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*, 29(8):1083–1085, April 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt090. URL `https://academic.oup.com/bioinformatics/article/29/8/1083/229442`. Publisher: Oxford Academic.

[29] Maayan Baron, Adrian Veres, SamuelL. Wolock, AubreyL. Faust, Renaud Gaujoux, Amedeo Vetere, JenniferHyoje Ryu, BridgetK. Wagner, ShaiS. Shen-Orr, AllonM. Klein, DouglasA. Melton, and Itai Yanai. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, 3(4):346–360.e4, October 2016. ISSN 2405-4712. doi: 10.1016/j.cels.2016.08.011. URL `http://www.sciencedirect.com/science/article/pii/S2405471216302666`.

[30] Jiebiao Wang, Bernie Devlin, and Kathryn Roeder. Using multiple measurements of tissue to estimate subject- and cell-type-specific gene expression. *Bioinformatics*, 36(3):782–788, February 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz619. URL `https://academic.oup.com/bioinformatics/article/36/3/782/5545976`. Publisher: Oxford Academic.

[31] Shai S. Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L. Bodian, Frank Staedtler, Nicholas M. Perry, Trevor Hastie, Minnie M. Sarwal, Mark M. Davis, and Atul J. Butte. Cell typespecific gene expression differences in complex tissues. *Nature Methods*, 7(4):287–289, April 2010. ISSN 1548-7105. doi: 10.1038/nmeth.1439. URL `https://www.nature.com/articles/nmeth.1439`. Number: 4 Publisher: Nature Publishing Group.

[32] Daifeng Wang, Shuang Liu, Jonathan Warrell, Hyejung Won, Xu Shi, Fabio C. P. Navarro, Declan Clarke, Mengting Gu, Prashant Emani, Yucheng T. Yang, Min Xu, Michael J. Gandal, Shaoke Lou, Jing Zhang, Jonathan J. Park, Chengfei Yan, Suhn Kyong Rhie, Kasidet Manakongtreecheep, Holly Zhou,

Aparna Nathan, Mette Peters, Eugenio Mattei, Dominic Fitzgerald, Tonya Brunetti, Jill Moore, Yan Jiang, Kiran Girdhar, Gabriel E. Hoffman, Selim Kalayci, Zeynep H. Gm, Gregory E. Crawford, PsychENCODE Consortium, Panos Roussos, Schahram Akbarian, Andrew E. Jaffe, Kevin P. White, Zhiping Weng, Nenad Sestan, Daniel H. Geschwind, James A. Knowles, and Mark B. Gerstein. Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362(6420), December 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aat8464. URL `https://science.sciencemag.org/content/362/6420/eaat8464`. Publisher: American Association for the Advancement of Science Section: Research Article.

[33] E. E. Burke, J. G. Chenoweth, J. H. Shin, L. Collado-Torres, S. K. Kim, N. Micali, Y. Wang, R. E. Straub, D. J. Hoeppner, H. Y. Chen, A. Lescure, K. Shibbani, G. R. Hamersky, B. N. Phan, W. S. Ulrich, C. Valencia, A. Jaishankar, A. J. Price, A. Rajpurohit, S. A. Semick, R. Brli, J. C. Barrow, D. J. Hiler, S. C. Page, K. Martinowich, T. M. Hyde, J. E. Kleinman, K. F. Berman, J. A. Apud, A. J. Cross, N. J. Brandon, D. R. Weinberger, B. J. Maher, R. D. G. McKay, and A. E. Jaffe. Dissecting transcriptomic signatures of neuronal differentiation and maturation using iPSCs. *bioRxiv*, page 380758, July 2018. doi: 10.1101/380758. URL `https://www.biorxiv.org/content/10.1101/380758v1`. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[34] Leonardo Collado-Torres, Emily E. Burke, Amy Peterson, JooHeon Shin, Richard E. Straub, Anandita Rajpurohit, Stephen A. Semick, William S.

Ulrich, Amanda J. Price, Cristian Valencia, Ran Tao, Amy Deep-Soboslay, Thomas M. Hyde, Joel E. Kleinman, Daniel R. Weinberger, and Andrew E. Jaffe. Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron*, 103(2):203–216.e8, July 2019. ISSN 0896-6273. doi: 10.1016/j.neuron.2019.05.013. URL `http://www.sciencedirect.com/science/article/pii/S0896627319304386`.

[35] Xiaoxiao Xu, Alan B. Wells, David R. O'Brien, Arye Nehorai, and Joseph D. Dougherty. Cell Type-Specific Expression Analysis to Identify Putative Cellular Mechanisms for Neurogenetic Disorders. *J. Neurosci.*, 34(4):1420–1431, January 2014. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4488-13.2014. URL `https://www.jneurosci.org/content/34/4/1420`. Publisher: Society for Neuroscience Section: Articles.

[36] Blue B. Lake, Rizi Ai, Gwendolyn E. Kaeser, Neeraj S. Salathia, Yun C. Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian-Bing Fan, Wei Wang, Jerold Chun, and Kun Zhang. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, 352(6293):1586–1590, June 2016. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaf1204. URL `https://science.sciencemag.org/content/352/6293/1586`. Publisher: American Association for the Advancement of Science Section: Reports.

[37] Blue B. Lake, Song Chen, Brandon C. Sos, Jean Fan, Gwendolyn E. Kaeser, Yun C. Yung, Thu E. Duong, Derek Gao, Jerold Chun, Peter V. Kharchenko, and Kun Zhang. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature Biotechnology*, 36(1):70–80, January 2018. ISSN 1546-1696. doi: 10.1038/nbt.4038. URL `https://www.nature.com/articles/nbt.4038`. Number: 1 Publisher: Nature Publishing Group.

[38] Dmitry Velmeshev, Lucas Schirmer, Diane Jung, Maximilian Haeussler, Yonatan Perez, Simone Mayer, Aparna Bhaduri, Nitasha Goyal, David H. Rowitch, and Arnold R. Kriegstein. Single-cell genomics identifies cell type-specific molecular changes in autism. *Science*, 364(6441):685–689, May 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aav8130. URL `https://science.sciencemag.org/content/364/6441/685`. Publisher: American Association for the Advancement of Science Section: Report.

[39] Rebecca D. Hodge, Trygve E. Bakken, Jeremy A. Miller, Kimberly A. Smith, Eliza R. Barkan, Lucas T. Graybuck, Jennie L. Close, Brian Long, Nelson Johansen, Osnat Penn, Zizhen Yao, Jeroen Eggermont, Thomas Hllt, Boaz P. Levi, Soraya I. Shehata, Brian Aevermann, Allison Beller, Darren Bertagnolli, Krissy Brouner, Tamara Casper, Charles Cobbs, Rachel Dalley, Nick Dee, Song-Lin Ding, Richard G. Ellenbogen, Olivia Fong, Emma Garren, Jeff Goldy, Ryder P. Gwinn, Daniel Hirschstein, C. Dirk Keene, Mohamed Keshk, Andrew L. Ko, Kanan Lathia, Ahmed Mahfouz, Zoe Maltzer, Medea McGraw, Thuc Nghi Nguyen, Julie Nyhus, Jeffrey G. Ojemann, Aaron Oldre,

Sheana Parry, Shannon Reynolds, Christine Rimorin, Nadiya V. Shapovalova, Saroja Somasundaram, Aaron Szafer, Elliot R. Thomsen, Michael Tieu, Gerald Quon, Richard H. Scheuermann, Rafael Yuste, Susan M. Sunkin, Boudewijn Lelieveldt, David Feng, Lydia Ng, Amy Bernard, Michael Hawrylycz, John W. Phillips, Bosiljka Tasic, Hongkui Zeng, Allan R. Jones, Christof Koch, and Ed S. Lein. Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 573(7772):61–68, September 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1506-7. URL `https://www.nature.com/articles/s41586-019-1506-7`. Number: 7772 Publisher: Nature Publishing Group.

[40] Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z. Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, Anthony J. Martorell, Richard M. Ransohoff, Brian P. Hafler, David A. Bennett, Manolis Kellis, and Li-Huei Tsai. Single-cell transcriptomic analysis of Alzheimers disease. *Nature*, 570(7761):332–337, June 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1195-2. URL `https://www.nature.com/articles/s41586-019-1195-2`. Number: 7761 Publisher: Nature Publishing Group.

[41] Christina A. Markunas, Stephen A. Semick, Bryan C. Quach, Ran Tao, Amy Deep-Soboslay, Laura J. Bierut, Thomas M. Hyde, Joel E. Kleinman, Eric O. Johnson, Andrew E. Jaffe, and Dana B. Hancock. Genome-wide DNA methylation differences in nucleus accumbens of smokers vs. nonsmokers. *bioRxiv*, page 781542, September 2019. doi: 10.1101/781542. URL

https://www.biorxiv.org/content/10.1101/781542v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[42] Andrew E. Jaffe, Jooheon Shin, Leonardo Collado-Torres, Jeffrey T. Leek, Ran Tao, Chao Li, Yuan Gao, Yankai Jia, Brady J. Maher, Thomas M. Hyde, Joel E. Kleinman, and Daniel R. Weinberger. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Neuroscience*, 18(1):154–161, January 2015. ISSN 1546-1726. doi: 10.1038/nn.3898. URL https://www.nature.com/articles/nn.3898. Number: 1 Publisher: Nature Publishing Group.

[43] Simone Codeluppi, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods*, 15(11): 932–935, November 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0175-z. URL https://www.nature.com/articles/s41592-018-0175-z. Number: 11 Publisher: Nature Publishing Group.

[44] Martin J. Aryee, Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, May 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu049. URL https://academic.oup.com/bioinformatics/article/30/10/1363/267584. Publisher: Oxford Academic.

[45] George F. Koob and Nora D. Volkow. Neurobiology of addiction: a neurocircuitry analysis. *Lancet Psychiatry*, 3(8):760–773, August 2016. ISSN 2215-0374. doi: 10.1016/S2215-0366(16)00104-8.

[46] Eric J. Nestler. Is there a common molecular pathway for addiction? *Nature Neuroscience*, 8(11):1445–1449, November 2005. ISSN 1546-1726. doi: 10.1038/nn1578. URL `https://www.nature.com/articles/nn1578`. Number: 11 Publisher: Nature Publishing Group.

[47] Andrew E. Jaffe, Richard E. Straub, Joo Heon Shin, Ran Tao, Yuan Gao, Leonardo Collado-Torres, Tony Kam-Thong, Hualin S. Xi, Jie Quan, Qiang Chen, Carlo Colantuoni, William S. Ulrich, Brady J. Maher, Amy Deep-Soboslay, Alan J. Cross, Nicholas J. Brandon, Jeffrey T. Leek, Thomas M. Hyde, Joel E. Kleinman, and Daniel R. Weinberger. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nature Neuroscience*, 21(8):1117–1125, August 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0197-y. URL `https://www.nature.com/articles/s41593-018-0197-y`. Number: 8 Publisher: Nature Publishing Group.

[48] Michael J. Gandal, Pan Zhang, Evi Hadjimichael, Rebecca L. Walker, Chao Chen, Shuang Liu, Hyejung Won, Harm van Bakel, Merina Varghese, Yongjun Wang, Annie W. Shieh, Jillian Haney, Sepideh Parhami, Judson Belmont, Minsoo Kim, Patricia Moran Losada, Zenab Khan, Justyna Mleczko, Yan Xia, Rujia Dai, Daifeng Wang, Yucheng T. Yang, Min Xu, Kenneth Fish, Patrick R. Hof, Jonathan Warrell, Dominic Fitzgerald, Kevin White, Andrew E. Jaffe,

PsychENCODE Consortium, Mette A. Peters, Mark Gerstein, Chunyu Liu, Lilia M. Iakoucheva, Dalila Pinto, and Daniel H. Geschwind. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*, 362(6420), December 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aat8127. URL https://science.sciencemag.org/content/362/6420/eaat8127. Publisher: American Association for the Advancement of Science Section: Research Article.

[49] Andrew E. Jaffe, Yuan Gao, Amy Deep-Soboslay, Ran Tao, Thomas M. Hyde, Daniel R. Weinberger, and Joel E. Kleinman. Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nature Neuroscience*, 19(1):40–47, January 2016. ISSN 1546-1726. doi: 10.1038/nn.4181. URL https://www.nature.com/articles/nn.4181. Number: 1 Publisher: Nature Publishing Group.

[50] Amanda J. Price, Taeyoung Hwang, Ran Tao, Emily E. Burke, Anandita Rajpurohi, Joo Heon Shin, Thomas M. Hyde, Joel E. Kleinman, Andrew E. Jaffe, and Daniel R. Weinberger. Characterizing the nuclear and cytoplasmic transcriptomes in developing and mature human cortex uncovers new insight into psychiatric disease gene regulation. *bioRxiv*, page 567966, March 2019. doi: 10.1101/567966. URL https://www.biorxiv.org/content/10.1101/567966v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[51] Trygve E. Bakken, Rebecca D. Hodge, Jeremy A. Miller, Zizhen Yao, Thuc Nghi Nguyen, Brian Aevermann, Eliza Barkan, Darren Bertagnolli,

Tamara Casper, Nick Dee, Emma Garren, Jeff Goldy, Lucas T. Graybuck, Matthew Kroll, Roger S. Lasken, Kanan Lathia, Sheana Parry, Christine Rimorin, Richard H. Scheuermann, Nicholas J. Schork, Soraya I. Shehata, Michael Tieu, John W. Phillips, Amy Bernard, Kimberly A. Smith, Hongkui Zeng, Ed S. Lein, and Bosiljka Tasic. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLOS ONE*, 13(12):e0209648, December 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0209648. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209648`. Publisher: Public Library of Science.

[52] Shijie C. Zheng, Charles E. Breeze, Stephan Beck, and Andrew E. Teschendorff. Identification of differentially methylated cell types in epigenome-wide association studies. *Nature Methods*, 15(12):1059–1066, December 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0213-x. URL `https://www.nature.com/articles/s41592-018-0213-x`. Number: 12 Publisher: Nature Publishing Group.

[53] Lars G. Fritsche, Robert N. Fariss, Dwight Stambolian, Gonalo R. Abecasis, Christine A. Curcio, and Anand Swaroop. Age-related macular degeneration: genetics and biology coming together. *Annu Rev Genomics Hum Genet*, 15: 151–171, 2014. ISSN 1545-293X. doi: 10.1146/annurev-genom-090413-025610.

[54] Lars G. Fritsche, Wilmar Igl, Jessica N. Cooke Bailey, Felix Grassmann, Sebanti Sengupta, Jennifer L. Bragg-Gresham, Kathryn P. Burdon, Scott J. Hebbring, Cindy Wen, Mathias Gorski, Ivana K. Kim, David Cho, Donald Zack,

Eric Souied, Hendrik P. N. Scholl, Elisa Bala, Kristine E. Lee, David J. Hunter, Rebecca J. Sardell, Paul Mitchell, Joanna E. Merriam, Valentina Cipriani, Joshua D. Hoffman, Tina Schick, Yara T. E. Lechanteur, Robyn H. Guymer, Matthew P. Johnson, Yingda Jiang, Chloe M. Stanton, Gabrille H. S. Buitendijk, Xiaowei Zhan, Alan M. Kwong, Alexis Boleda, Matthew Brooks, Linn Gieser, Rinki Ratnapriya, Kari E. Branham, Johanna R. Foerster, John R. Heckenlively, Mohammad I. Othman, Brendan J. Vote, Helena Hai Liang, Emmanuelle Souzeau, Ian L. McAllister, Timothy Isaacs, Janette Hall, Stewart Lake, David A. Mackey, Ian J. Constable, Jamie E. Craig, Terrie E. Kitchner, Zhenglin Yang, Zhiguang Su, Hongrong Luo, Daniel Chen, Hong Ouyang, Ken Flagg, Danni Lin, Guanping Mao, Henry Ferreyra, Klaus Stark, Claudia N. von Strachwitz, Armin Wolf, Caroline Brandl, Guenther Rudolph, Matthias Olden, Margaux A. Morrison, Denise J. Morgan, Matthew Schu, Jeeyun Ahn, Giuliana Silvestri, Evangelia E. Tsironi, Kyu Hyung Park, Lindsay A. Farrer, Anton Orlin, Alexander Brucker, Mingyao Li, Christine A. Curcio, Saddek Mohand-Sad, Jos-Alain Sahel, Isabelle Audo, Mustapha Benchaboune, Angela J. Cree, Christina A. Rennie, Srinivas V. Goverdhan, Michelle Grunin, Shira Hagbi-Levi, Peter Campochiaro, Nicholas Katsanis, Frank G. Holz, Frdric Blond, Hlne Blanch, Jean-Franois Deleuze, Robert P. Igo, Barbara Truitt, Neal S. Peachey, Stacy M. Meuer, Chelsea E. Myers, Emily L. Moore, Ronald Klein, Michael A. Hauser, Eric A. Postel, Monique D. Courtenay, Stephen G. Schwartz, Jaclyn L. Kovach, William K. Scott, Gerald Liew, Ava G. Tan, Bamini Gopinath, John C. Merriam, R. Theodore Smith, Jane C. Khan,

Humma Shahid, Anthony T. Moore, J. Allie McGrath, Rene Laux, Milam A. Brantley, Anita Agarwal, Lebriz Ersoy, Albert Caramoy, Thomas Langmann, Nicole T. M. Saksens, Eiko K. de Jong, Carel B. Hoyng, Melinda S. Cain, Andrea J. Richardson, Tammy M. Martin, John Blangero, Daniel E. Weeks, Bal Dhillon, Cornelia M. van Duijn, Kimberly F. Doheny, Jane Romm, Caroline C. W. Klaver, Caroline Hayward, Michael B. Gorin, Michael L. Klein, Paul N. Baird, Anneke I. den Hollander, Sascha Fauser, John R. W. Yates, Rando Allikmets, Jie Jin Wang, Debra A. Schaumberg, Barbara E. K. Klein, Stephanie A. Hagstrom, Itay Chowers, Andrew J. Lotery, Thierry Lveillard, Kang Zhang, Murray H. Brilliant, Alex W. Hewitt, Anand Swaroop, Emily Y. Chew, Margaret A. Pericak-Vance, Margaret DeAngelis, Dwight Stambolian, Jonathan L. Haines, Sudha K. Iyengar, Bernhard H. F. Weber, Gonalo R. Abecasis, and Iris M. Heid. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.*, 48(2):134–143, February 2016. ISSN 1546-1718. doi: 10.1038/ng.3448.

[55] Timothy W. Olsen and Xiao Feng. The Minnesota Grading System of Eye Bank Eyes for Age-Related Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.*, 45(12):4484–4490, December 2004. ISSN 1552-5783. doi: 10.1167/iovs.04-0342. URL `https://iovs.arvojournals.org/article.aspx?articleid=2163244`. Publisher: The Association for Research in Vision and Ophthalmology.

[56] Frederick L. Ferris, Matthew D. Davis, Traci E. Clemons, Li-Yin Lee, Emily Y. Chew, Anne S. Lindblad, Roy C. Milton, Susan B. Bressler, Ronald Klein,

and Age-Related Eye Disease Study (AREDS) Research Group. A simplified severity scale for age-related macular degeneration: AREDS Report No. 18. *Arch. Ophthalmol.*, 123(11):1570–1574, November 2005. ISSN 0003-9950. doi: 10.1001/archopht.123.11.1570.

[57] ALEJANDRA DECANINI, CURTIS L. NORDGAARD, XIAO FENG, DEBORAH A. FERRINGTON, and TIMOTHY W. OLSEN. Changes in Select Redox Proteins of the Retinal Pigment Epithelium in Age-related Macular Degeneration. *Am J Ophthalmol*, 143(4):607–615, April 2007. ISSN 0002-9394. doi: 10.1016/j.ajo.2006.12.006. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2365890/`.

[58] Johann A. Gagnon-Bartsch and Terence P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, July 2012. ISSN 1465-4644. doi: 10.1093/biostatistics/kxr034. URL `https://academic.oup.com/biostatistics/article/13/3/539/248166`. Publisher: Oxford Academic.

[59] Jeffrey T. Leek. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.*, 42(21), December 2014. ISSN 1362-4962. doi: 10.1093/nar/gku864.

[60] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, March 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts034.

[61] Eli Eisenberg and Erez Y. Levanon. Human housekeeping genes, revisited. *Trends Genet.*, 29(10):569–574, October 2013. ISSN 0168-9525. doi: 10.1016/j.tig.2013.05.010.

[62] Andreas Scherer. *Batch Effects and Noise in Microarray Experiments*. John Wiley & Sons, Ltd, 1 edition, 2009. doi: 10.1002/9780470685983. URL https://onlinelibrary.wiley.com/doi/10.1002/9780470685983. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470685983.

[63] Marta Mel, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, Jakob M. Goldmann, Dmitri D. Pervouchine, Timothy J. Sullivan, Rory Johnson, Ayellet V. Segr, Sarah Djebali, Anastasia Niarchou, The GTEx Consortium, Fred A. Wright, Tuuli Lappalainen, Miquel Calvo, Gad Getz, Emmanouil T. Dermitzakis, Kristin G. Ardlie, and Roderic Guig. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, May 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa0355. URL https://science.sciencemag.org/content/348/6235/660. Publisher: American Association for the Advancement of Science Section: Report.

[64] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool

for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556. URL https://www.nature.com/articles/ng0500_25. Number: 1 Publisher: Nature Publishing Group.

[65] The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res*, 45(D1):D331–D338, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1108. URL https://academic.oup.com/nar/article/45/D1/D331/2605810. Publisher: Oxford Academic.

[66] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, 16(5):284–287, May 2012. ISSN 1557-8100. doi: 10.1089/omi.2011.0118.

[67] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, May 2010. ISSN 1546-1696. doi: 10.1038/nbt.1621. URL https://www.nature.com/articles/nbt.1621. Number: 5 Publisher: Nature Publishing Group.

[68] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17):2325–2329, September 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr355. URL https://academic.oup.com/bioinformatics/article/27/17/2325/223194. Publisher: Oxford Academic.

[69] Olivier Delaneau, Halit Ongen, Andrew A. Brown, Alexandre Fort, Nikolaos I. Panousis, and Emmanouil T. Dermitzakis. A complete tool set for molecular QTL discovery and analysis. *Nature Communications*, 8(1):1–7, May 2017. ISSN 2041-1723. doi: 10.1038/ncomms15452. URL `https://www.nature.com/articles/ncomms15452`. Number: 1 Publisher: Nature Publishing Group.

[70] Nick Patterson, Alkes L. Price, and David Reich. Population Structure and Eigenanalysis. *PLOS Genetics*, 2(12):e190, December 2006. ISSN 1553-7404. doi: 10.1371/journal.pgen.0020190. URL `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020190`. Publisher: Public Library of Science.

[71] Andrey A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, May 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts163. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3348564/`.

[72] T. Mark Beasley, Stephen Erickson, and David B. Allison. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav. Genet.*, 39(5):580–595, September 2009. ISSN 1573-3297. doi: 10.1007/s10519-009-9281-0.

[73] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, August 2006. ISSN 1546-1718. doi: 10.1038/ng1847. URL `https://`

www.nature.com/articles/ng1847. Number: 8 Publisher: Nature Publishing Group.

[74] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr330. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/.

[75] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W. J. H. Penninx, Rick Jansen, Eco J. C. de Geus, Dorret I. Boomsma, Fred A. Wright, Patrick F. Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J. Lusis, Terho Lehtimki, Emma Raitoharju, Mika Khnen, Ilkka Seppl, Olli T. Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L. Price, Pivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, March 2016. ISSN 1546-1718. doi: 10.1038/ng.3506. URL https://www.nature.com/articles/ng.3506. Number: 3 Publisher: Nature Publishing Group.

[76] Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, 88(1): 76–82, January 2011. ISSN 1537-6605. doi: 10.1016/j.ajhg.2010.11.011.

[77] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law,

Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47, April 2015. ISSN 1362-4962. doi: 10.1093/nar/gkv007.

[78] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0506580102. URL `http://www.pnas.org/cgi/doi/10.1073/pnas.0506580102`.

[79] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, December 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-559. URL `https://doi.org/10.1186/1471-2105-9-559`.

[80] Aaron M. Newman, Natasha B. Gallo, Lisa S. Hancox, Norma J. Miller, Carolyn M. Radeke, Michelle A. Maloney, James B. Cooper, Gregory S. Hageman, Don H. Anderson, Lincoln V. Johnson, and Monte J. Radeke. Systems-level analysis of age-related macular degeneration reveals global biomarkers and phenotype-specific functional networks. *Genome Med*, 4(2):16, February 2012. ISSN 1756-994X. doi: 10.1186/gm315.

[81] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker.

Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, November 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303.

[82] Michele Pinelli, Annamaria Carissimo, Luisa Cutillo, Ching-Hung Lai, Margherita Mutarelli, Maria Nicoletta Moretti, Marwah Veer Singh, Marianthi Karali, Diego Carrella, Mariateresa Pizzo, Francesco Russo, Stefano Ferrari, Diego Ponzin, Claudia Angelini, Sandro Banfi, and Diego di Bernardo. An atlas of gene expression and gene co-regulation in the human retina. *Nucleic Acids Res.*, 44(12):5773–5784, 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw486.

[83] Q. V. Hoang, R. A. Linsenmeier, C. K. Chung, and C. A. Curcio. Photoreceptor inner segments in monkey and human retina: mitochondrial density, optics, and regional variation. *Vis. Neurosci.*, 19(4):395–407, August 2002. ISSN 0952-5238. doi: 10.1017/s0952523802194028.

[84] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson. Human photoreceptor topography. *J. Comp. Neurol.*, 292(4):497–523, February 1990. ISSN 0021-9967. doi: 10.1002/cne.902920402.

[85] Hilary K. Finucane, Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, Giulio Genovese, Arpiar Saunders, Evan Macosko, Samuela Pollack, John R. B. Perry, Jason D. Buenrostro, Bradley E. Bernstein, Soumya Raychaudhuri, Steven McCarroll, Benjamin M. Neale, and Alkes L. Price. Heritability enrichment of specifically expressed genes identifies disease-relevant

tissues and cell types. *Nature Genetics*, 50(4):621–629, April 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0081-4. URL https://www.nature.com/articles/s41588-018-0081-4. Number: 4 Publisher: Nature Publishing Group.

[86] Eric R. Gamazon, Ayellet V. Segr, Martijn van de Bunt, Xiaoquan Wen, Hualin S. Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M. Derks, Franois Aguet, Jie Quan, Dan L. Nicolae, Eleazar Eskin, Manolis Kellis, Gad Getz, Mark I. McCarthy, Emmanouil T. Dermitzakis, Nancy J. Cox, and Kristin G. Ardlie. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics*, 50(7):956–967, July 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0154-4. URL https://www.nature.com/articles/s41588-018-0154-4. Number: 7 Publisher: Nature Publishing Group.

[87] Tobias Strunz, Felix Grassmann, Javier Gayn, Satu Nahkuri, Debora Souza-Costa, Cyrille Maugeais, Sascha Fauser, Everson Nogoceke, and Bernhard H. F. Weber. A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci Rep*, 8(1):5865, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24219-z.

[88] Menachem Fromer, Panos Roussos, Solveig K. Sieberts, Jessica S. Johnson, David H. Kavanagh, Thanneer M. Perumal, Douglas M. Ruderfer, Edwin C. Oh, Aaron Topol, Hardik R. Shah, Lambertus L. Klei, Robin Kramer, Dalila Pinto, Zeynep H. Gm, A. Ercument Cicek, Kristen K. Dang, Andrew Browne,

Cong Lu, Lu Xie, Ben Readhead, Eli A. Stahl, Jianqiu Xiao, Mahsa Parvizi, Tymor Hamamsy, John F. Fullard, Ying-Chih Wang, Milind C. Mahajan, Jonathan M. J. Derry, Joel T. Dudley, Scott E. Hemby, Benjamin A. Logsdon, Konrad Talbot, Towfique Raj, David A. Bennett, Philip L. De Jager, Jun Zhu, Bin Zhang, Patrick F. Sullivan, Andrew Chess, Shaun M. Purcell, Leslie A. Shinobu, Lara M. Mangravite, Hiroyoshi Toyoshiba, Raquel E. Gur, Chang-Gyu Hahn, David A. Lewis, Vahram Haroutunian, Mette A. Peters, Barbara K. Lipska, Joseph D. Buxbaum, Eric E. Schadt, Keisuke Hirai, Kathryn Roeder, Kristen J. Brennand, Nicholas Katsanis, Enrico Domenici, Bernie Devlin, and Pamela Sklar. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.*, 19(11):1442–1453, 2016. ISSN 1546-1726. doi: 10.1038/nn.4399.

[89] Michael J. Gandal, Jillian R. Haney, Neelroop N. Parikshak, Virpi Leppa, Gokul Ramaswami, Chris Hartl, Andrew J. Schork, Vivek Appadurai, Alfonso Buil, Thomas M. Werge, Chunyu Liu, Kevin P. White, CommonMind Consortium, PsychENCODE Consortium, iPSYCH-BROAD Working Group, Steve Horvath, and Daniel H. Geschwind. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science*, 359(6376): 693–697, February 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aad6469. URL `https://science.sciencemag.org/content/359/6376/693`. Publisher: American Association for the Advancement of Science Section: Report.

[90] Tim Beck, Robert K. Hastings, Sirisha Gollapudi, Robert C. Free, and Anthony J. Brookes. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet.*, 22(7):949–952, July 2014. ISSN 1476-5438. doi: 10.1038/ejhg.2013.274.

[91] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*, 45(Database issue):D896–D901, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1133. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210590/`.

[92] Aravinda Chakravarti, Andrew G. Clark, and Vamsi K. Mootha. Distilling Pathophysiology from Complex Disease Genetics. *Cell*, 155(1):21–26, September 2013. ISSN 0092-8674. doi: 10.1016/j.cell.2013.09.001. URL `http://www.sciencedirect.com/science/article/pii/S0092867413010921`.

[93] Michael D. Gallagher and Alice S. Chen-Plotkin. The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics*, 102(5):717–730, May 2018. ISSN 0002-9297. doi: 10.1016/j.ajhg.2018.04.002. URL `http://www.sciencedirect.com/science/article/pii/S0002929718301344`.

[94] Christopher D. Brown, Lara M. Mangravite, and Barbara E. Engelhardt. Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLOS Genetics*, 9 (8):e1003649, August 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003649. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003649. Publisher: Public Library of Science.

[95] Krisztina Kozma, Jeremy J. Keusch, Bjrn Hegemann, Kelvin B. Luther, Dominique Klein, Daniel Hess, Robert S. Haltiwanger, and Jan Hofsteenge. Identification and Characterization of a1,3-Glucosyltransferase That Synthesizes the Glc-1,3-Fuc Disaccharide on Thrombospondin Type 1 Repeats. *J. Biol. Chem.*, 281(48):36742–36751, December 2006. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M605912200. URL http://www.jbc.org/content/281/48/36742. Publisher: American Society for Biochemistry and Molecular Biology.

[96] Saskia A. J. LesnikOberstein, Marjolein Kriek, Stefan J. White, Margot E. Kalf, Karoly Szuhai, Johan T. denDunnen, Martijn H. Breuning, and Raoul C. M. Hennekam. Peters Plus Syndrome Is Caused by Mutations in B3GALTL, a Putative Glycosyltransferase. *Am J Hum Genet*, 79(3):562–566, September 2006. ISSN 0002-9297. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1559553/.

[97] Lars Langemeyer and Christian Ungermann. BORC and BLOC-1: Shared Subunits in Trafficking Complexes. *Developmental Cell*, 33(2):121–122, April

2015. ISSN 1534-5807. doi: 10.1016/j.devcel.2015.04.008. URL http://www.sciencedirect.com/science/article/pii/S1534580715002506.

[98] Ariana P. Mullin, Madhumala K. Sadanandappa, Wenpei Ma, Dion K. Dickman, Krishnaswamy VijayRaghavan, Mani Ramaswami, Subhabrata Sanyal, and Victor Faundez. Gene dosage in the dysbindin schizophrenia susceptibility network differentially affect synaptic function and plasticity. *J. Neurosci.*, 35(1):325–338, January 2015. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.3542-14.2015.

[99] Farhad Hormozdiari, Steven Gazal, Bryce van de Geijn, Hilary K. Finucane, Chelsea J.-T. Ju, Po-Ru Loh, Armin Schoech, Yakir Reshef, Xuanyao Liu, Luke OConnor, Alexander Gusev, Eleazar Eskin, and Alkes L. Price. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature Genetics*, 50(7):1041–1047, July 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0148-2. URL https://www.nature.com/articles/s41588-018-0148-2. Number: 7 Publisher: Nature Publishing Group.

[100] Xuanyao Liu, Yang I Li, and Jonathan K Pritchard. Trans effects on gene expression can drive omnigenic inheritance. preprint, Genetics, September 2018. URL http://biorxiv.org/lookup/doi/10.1101/425108.

[101] Luke R. Lloyd-Jones, Alexander Holloway, Allan McRae, Jian Yang, Kerrin Small, Jing Zhao, Biao Zeng, Andrew Bakshi, Andres Metspalu, Manolis Dermitzakis, Greg Gibson, Tim Spector, Grant Montgomery, Tonu Esko,

Peter M. Visscher, and Joseph E. Powell. The Genetic Architecture of Gene Expression in Peripheral Blood. *The American Journal of Human Genetics*, 100(2):228–237, February 2017. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2016.12.008. URL https://www.cell.com/ajhg/abstract/S0002-9297(16)30532-8.

[102] Klaasjan G. Ouwens, Rick Jansen, Michel G. Nivard, Jenny van Dongen, Maia J. Frieser, Jouke-Jan Hottenga, Wibowo Arindrarto, Annique Claringbould, Maarten van Iterson, Hailiang Mei, Lude Franke, Bastiaan T. Heijmans, Peter A. C. 't Hoen, Joyce van Meurs, Andrew I. Brooks, Brenda W. J. H. Penninx, and Dorret I. Boomsma. A characterization of cis - and trans -heritability of RNA-Seq-based gene expression. *European Journal of Human Genetics*, pages 1–11, September 2019. ISSN 1476-5438. doi: 10.1038/s41431-019-0511-5. URL https://www.nature.com/articles/s41431-019-0511-5.

[103] Alkes L. Price, Agnar Helgason, Gudmar Thorleifsson, Steven A. McCarroll, Augustine Kong, and Kari Stefansson. Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals. *PLOS Genetics*, 7(2):e1001317, February 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001317. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001317.

[104] Ana Vinuela, Andrew A. Brown, Alfonso Buil, Pei-Chien Tsai, Matthew N. Davies, Jordana T. Bell, Emmanouil T. Dermitzakis, Timothy D. Spector, and Kerrin S. Small. Age-dependent changes in mean and variance of gene

expression across tissues in a twin cohort. *Hum Mol Genet*, 27(4):732–741, February 2018. ISSN 0964-6906. doi: 10.1093/hmg/ddx424. URL https://academic.oup.com/hmg/article/27/4/732/4710066.

[105] Yan Zhang, Guanghao Qi, Ju-Hyun Park, and Nilanjan Chatterjee. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature Genetics*, 50(9):1318, September 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0193-x. URL https://www.nature.com/articles/s41588-018-0193-x.

[106] Rainer Storn. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, page 19, 1997.

[107] T. Lumley and P. Heagerty. Weighted empirical adaptive variance estimators for correlated data regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):459–477, 1999. ISSN 1369-7412, 1467-9868. doi: 10.1111/1467-9868.00187. URL http://doi.wiley.com/10.1111/1467-9868.00187.

[108] Luke J. O'Connor, Armin P. Schoech, Farhad Hormozdiari, Steven Gazal, Nick Patterson, and Alkes L. Price. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *The American Journal of Human Genetics*, 105(3):456–476, September 2019. ISSN 0002-9297. doi: 10.1016/j.ajhg.2019.07.003. URL http://www.sciencedirect.com/science/article/pii/S0002929719302666.

[109] Joseph E. Powell, Anjali K. Henders, Allan F. McRae, Jinhee Kim, Gibran Hemani, Nicholas G. Martin, Emmanouil T. Dermitzakis, Greg Gibson, Grant W. Montgomery, and Peter M. Visscher. Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. *PLOS Genetics*, 9(5):e1003502, May 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003502. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003502. Publisher: Public Library of Science.

[110] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLOS ONE*, 9(1):e78644, January 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0078644. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078644. Publisher: Public Library of Science.

[111] Doug Speed, Na Cai, the UCLEB Consortium, Michael R. Johnson, Sergey Nejentsev, and David J. Balding. Reevaluation of SNP heritability in complex human traits. *Nature Genetics*, 49(7):986–992, July 2017. ISSN 1546-1718. doi: 10.1038/ng.3865. URL https://www.nature.com/articles/ng.3865.

[112] Luz D. Orozco, Hsu-Hsin Chen, Christian Cox, Kenneth J. Katschke, Rommel Arceo, Carmina Espiritu, Patrick Caplazi, Sarajane Saturnio Nghiem, Ying-Jiun Chen, Zora Modrusan, Amy Dressen, Leonard D. Goldstein, Christine Clarke, Tushar Bhangale, Brian Yaspan, Marion Jeanne, Michael J.

Townsend, Menno van Lookeren Campagne, and Jason A. Hackney. Integration of eQTL and a Single-Cell Atlas in the Human Eye Identifies Causal Genes for Age-Related Macular Degeneration. *Cell Reports*, 30(4):1246–1259.e6, January 2020. ISSN 2211-1247. doi: 10.1016/j.celrep.2019.12.082. URL `http://www.sciencedirect.com/science/article/pii/S2211124719317474`.

[113] Fei-Fei Cheng, You-Yuan Zhuang, Xin-Ran Wen, Angli Xue, Jian Yang, and Zi-Bing Jin. Towards the identification of causal genes for age-related macular degeneration. *bioRxiv*, page 778613, September 2019. doi: 10.1101/778613. URL `https://www.biorxiv.org/content/10.1101/778613v1`. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Kayode A Sosina
1115 Dlong road Apt C Catonsville, MD 21228. Email: kayodesosina@gmail.com Cell: 617-501-4426

EDUCATION Ph.D. Biostatistics, Johns Hopkins University, expected 05/2020; M.Sc. Biostatistics Harvard University; B.Sc. Mathematics and Statistics, University of Lagos, Nigeria.

EXPERIENCE

Johns Hopkins School of Public Health *Graduate research* 2017-Present

- Developed a method to model the effect size distribution across gene-regulatory trans-variants to better understand genetic architecture of gene expression.
- Integrated both gene expression and Genome-wide association studies (GWAs) data across multiple tissues, using summary level information, to better identify tissues driving the association signals.
- Worked on a project to identify strategies to address bias observed in cell-type proportion estimation using sc-RNA seq data.

National Eye Institute *Pre-doctoral visiting fellow* 2016-2018

- Lead research investigating the genetic contribution of gene expression on phenotypic variation in a population of individuals experiencing advanced macular degeneration.

Brigham and Women's Hospital *Research Assistant* 2014-2015

- Implemented models to investigate the effect of medication reconciliation on emergency department visits and hospitalizations in patients with diabetes using electronic health records (EHR).
- Worked on the validation of a Personalized, Patient-Centered Glycemic Control Benchmarking Tool for Type 2 Diabetes Mellitus (T2DM).

SKILLS

Highly skilled and proficient in R, SQL, SAS, STATA
Proficient in C++, Python

LEADERSHIP

Johns Hopkins School of Public Health

MPH Capstone Mentor	2019
Teaching Assistant, Analysis of Longitudinal Data	2017-2019
Teaching Assistant, Multilevel Statistical Models in Public Health	2017-2019
Teaching Assistant, Statistical Methods in Public Health	2016-2018
Teaching Assistant, Intro. Stats for Med. Research	2016

Harvard University

Teaching Fellow, Intro. Stats for Med. Research	2016

AWARDS

The David B. Duncan Centennial Scholarship (Johns Hopkins)	2018
Presidential Scholars Fund (Harvard University)	2013 - 2015
Endowment Fund for Honors Students (University of Lagos)	2009 – 2011

PUBLICATIONS(* co-first)

Published

Ratnapriya, R*. **Sosina, O**.*, Starostik, M.*, Kwacklis, M., Kapphahn, R., Walton, A., Pietraszkiewicz, A., Montezuma, S., Fritsche, L., Chew, E., Abecasis, G., Ferrington, D., Chatterjee, N., and Swarooop, A (2019). "Retinal transcriptome and eQTL analyses identify genes for age-related macular degeneration". Nature Genetics. https://doi.org/10.1038/s41588-019-0351-9

Turchin, A., **Sosina, O**., Zhang, H., Shubina, M., Desai, S. P., Simonson, D. C., & Testa, M.A.(2018). Ambulatory Medication Reconciliation and Frequency of Hospitalizations and Emergency Department Visits in Patients with Diabetes. Diabetes Care, 41(8), 1639-1645. https://doi.org/10.2337/dc17-1260.

Testa, M. A, Turchin, A., **Sosina, O**., & Simonson, D., (2015). Benchmarking T2DM Treatment Effectiveness in Clinical Practice. Poster presented at: 75th Scientific Session of the American Diabetes Association; 2015 June 9-13; Boston, MA.

Not Peer-Reviewed

Sosina Olukayode, Matthew N Tran, Kristen R Maynard, Ran Tao, Margaret Taub, Keri Martinowich, Steven A. Semick, Thomas M. Hyde, Dana B. Hancock, Joel E. Kleinman, Jeffrey T Leek, Andrew E Jaffe. "Strategies for cellular deconvolution in human brain RNA sequencing data". bioRxiv. https://doi.org/10.1101/2020.01.19.910976

Margaret R. Starostik, Olukayode A. Sosina, Rajiv C. McCoy. “Single-cell analysis of human embryos reveals diverse patterns of aneuploidy and mosaicism”. bioRxiv. https://doi.org/10.1101/2020.01.06.894287

In preparation

Sosina Olukayode, Nilanjan Chatterjee. “Estimation of effect size distribution for trans regulatory variants”.

Andrew Skol, Segun Jung, Ana Marija Sokovic, Siquan Chen, Sarah Fazal, Olukayode Sosina, Amy Lin, Maria Sverdlov, Poulami Borkar, Dingcai Cao, Anand Swaroop, Ionut Bebu, DCCT/ EDIC Study group, Barbara Stranger, Michael A. Grassi. “Mendelian randomization identifies FLCN expression as a mediator of diabetic retinopathy”.

Appendix A

Chapter 2

A.1 Sample processing and data generation for NAc (nucleus accumbens)

Single-nucleus RNA-seq data generation and processing in NAc

We performed single-nucleus RNA-seq (snRNA-seq) on nucleus accumbens (NAc) tissue from two donors using 10x Genomics Single Cell Gene Expression V3 technology. Nuclei were isolated using a Frankenstein nuclei isolation protocol developed by Martelotto et al. for frozen tissues. Briefly, ∼40mg of frozen NAc tissue was homogenized in chilled Nuclei EZ Lysis Buffer (MilliporeSigma) in a glass dounce with ∼15 strokes per pestle. Homogenate was filtered using a $70\mu m$-strainer mesh and centrifuged at 500xg for 5 minutes at 4°C in a benchtop centrifuge. Nuclei were resuspended in the EZ lysis buffer, centrifuged again, and equilibrated to nuclei wash/resuspension buffer (1x PBS, 1% BSA, 0.2U/uL RNase Inhibitor). Nuclei were

washed and centrifuged in this nuclei wash/resuspension buffer three times, before labeling with DAPI (10μg/mL). The sample was then filtered through a 35μm-cell strainer and sorted on a BD FACS Aria II Flow Cytometer (Becton Dickinson) at the Johns Hopkins University Sidney Kimmel Comprehensive Cancer Center (SKCCC) Flow Cytometry Core. Gating criteria hierarchically selected for whole, singlet nuclei (by forward/side scatter), then for G0/G1 nuclei (by DAPI fluorescence). A null sort was additionally performed from the same preparation to ensure nuclei input was free of debris. Approximately 8,500 single nuclei were sorted directly into 25.1μL of reverse transcription reagents from the 10x Genomics Single Cell 3 Reagents kit (without enzyme). Libraries were prepared according to manufacturers instructions (10x Genomics) and sequenced on the Next-seq (Illumina) at the Johns Hopkins University Transcriptomics and Deep Sequencing Core.

We processed the sequencing data with the 10x Genomics Cell Ranger pipeline, aligning to the human reference genome GRCh38, with a reconfigured GTF such that intronic alignments were additionally counted given the nuclear context, to generate UMI/feature-barcode matrices. We used R package Seurat for raw feature-barcode quality control, dimensionality reduction (PCA), choosing the top 30 PCs as the optimal dimensions for clustering. We performed graph-based clustering with the default Louvain approach, taking a computed K-nearest neighbors graph as input, which were then annotated with well-established cell type markers for nuclear type identity. We also used Seurats implementation of non-linear dimensionality reduction techniques, t-SNE and UMAP, simply for visualization of the high-dimensional structure in the data, which complemented the clustering results (Figure A.4). With

the five broad cell type annotations (neurons, oligodendrocytes, oligodendrocyte precursors, astrocytes, and microglia) of nuclear clusters, we identified unbiased cluster-driving genes (with Seurats FindAllMarkers() function, using the Wilcoxon rank-sum test), that were upregulated in each cell type/cluster, compared to all other nuclei. Using the same set of 24,048 genes, we have 4,169 high-quality nuclei in this reference, evenly distributed across donors. The top 50- and top 25-per-cell-type gene sets had 247 and 125 genes, respectively, which included many cell type marker genes used for annotation.

Bulk NAc Data Generation and Processing

Briefly, the nucleus accumbens (NAc) was dissected under visual guidance using a hand-held dental drill. Samples were obtained from the ventral striatum, anterior to the optic chiasm, at the level where the NAcforms a bridge between the putamen and the head of the caudate. DNA and RNA were concurrently extracted from dissected tissue using the Qiagen AllPrep DNA/RNA Mini Kit (Cat No./ID: 80204).

NAc DNA was profiled with the Infinium MethylationEPIC microarray using the manufacturer's protocol. Raw idat files were processed and normalized using the minfi Bioconductor package using stratified quantile normalization. Resulting neuronal fractions were estimated using the minfi estimateCellCounts function using sorted reference data from the DLPFC for neurons and non-neurons using the Houseman algorithm.

NAc RNA was subjected to RNA-seq library preparations using the Illumina RiboZero Gold kits and sequenced using 2x100bp paired end reads on an Illumina

HiSeq 3000.

A.2 Supplementary Figures

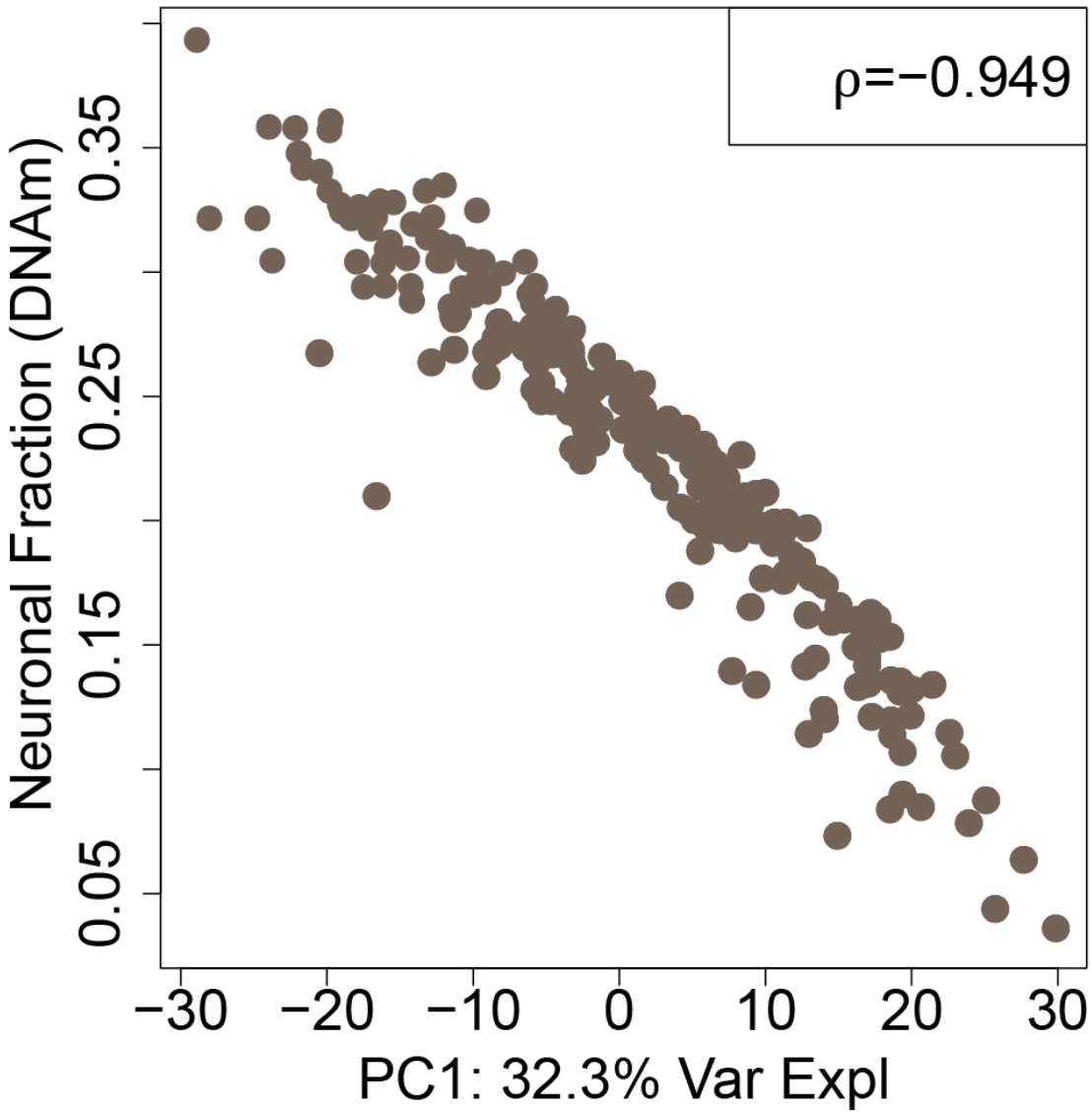


Fig. A.1. DNAm estimated neuronal fractions vs PC1. Scatter plot of neuronal fractions estimated using the Houseman approach with a DNAm reference vs the first principal component estimated from the bulk RNA-Seq data.

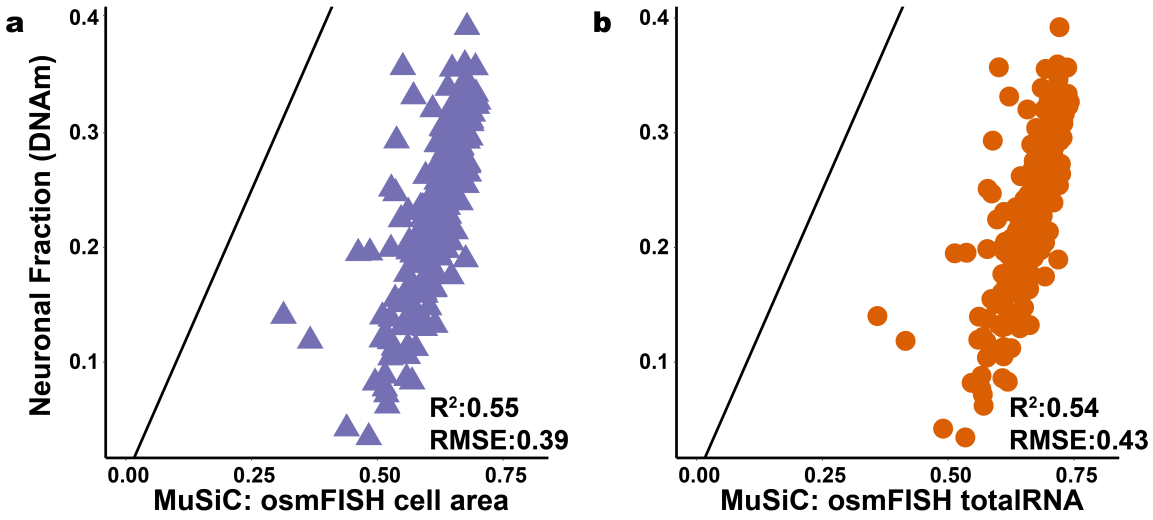


Fig. A.2. Deconvolution in bulk NAc data using gene expression profiles from the temporal cortex (Darmanis et al) with cell size estimates derived using mouse samples (osmFISH estimates of cell size). Scatter plots comparing neuronal fraction estimated for each individual using DNAm data and the Houseman method vs neuronal fractions based on scRNA-seq data and estimated using MuSiC with (**a**) osmFISH cell area as cell size, and (**b**) osmFISH total RNA molecule count as cell size.

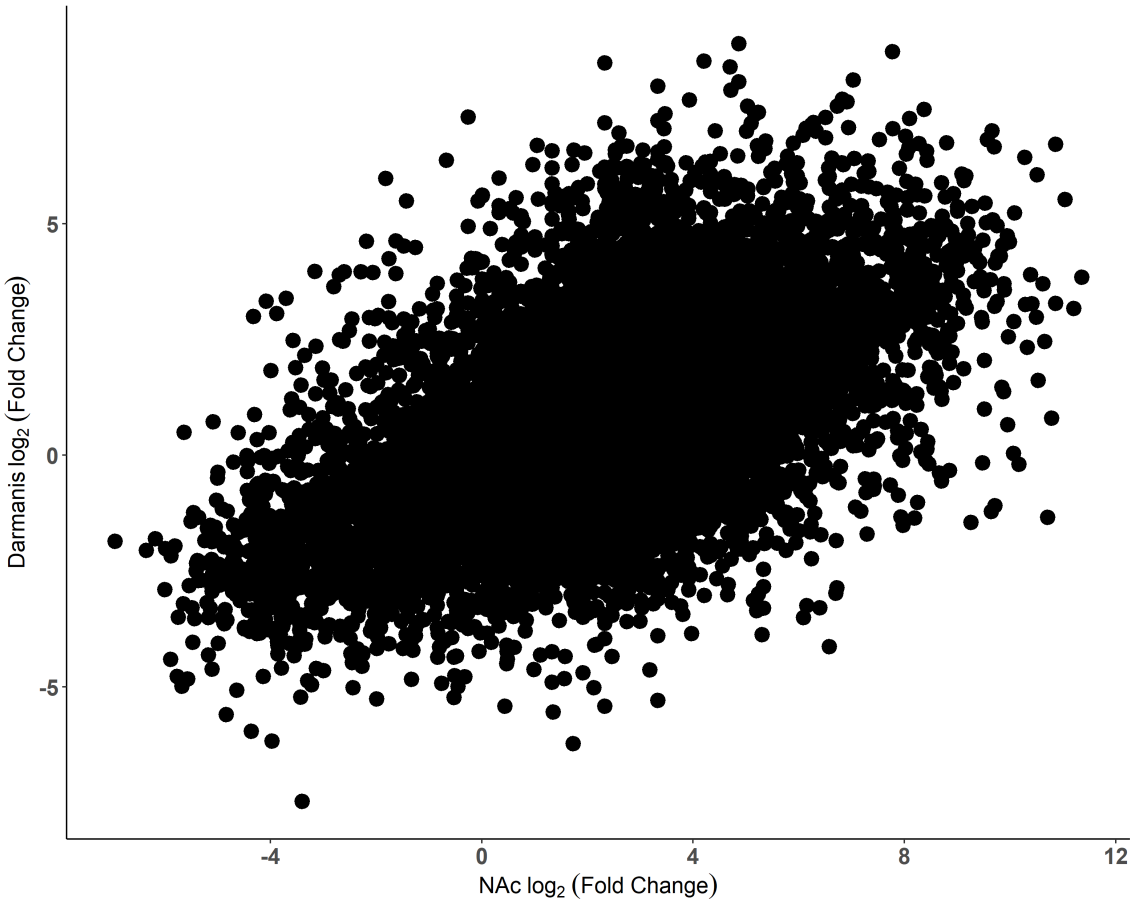


Fig. A.3. Neuronal enrichment of gene expression in scRNA-seq from temporal cortex and snRNA-seq from nucleus accumbens. Scatter plot shows the relationship, based on $\log_2$(fold change) comparing neuronal to glial, between the Darmanis reference dataset (y-axis) and the NAc reference dataset (x-axis). Each dot represents an estimated $\log_2$(fold change) for a given gene.

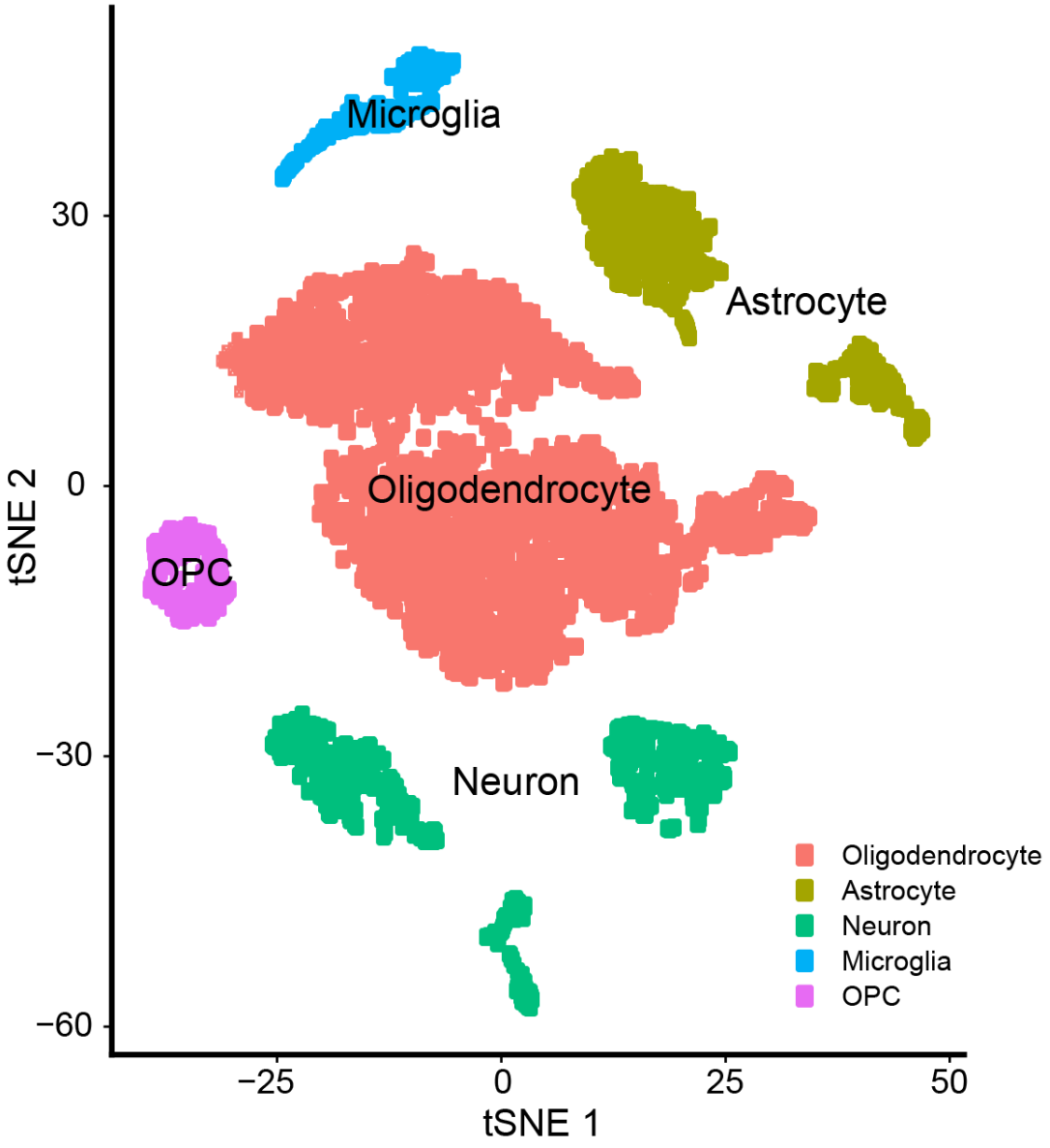


Fig. A.4. t-distributed stochastic neighbor embedding (t-SNE) of single-nucleus RNA-seq data from the two postmortem NAc samples, representing the 4,169 high-quality nuclei after processing. Nuclei are colored by cell type annotation after graph-based clustering, which shown here is largely in agreement with t-SNE coordinates. OPC represents Oligodendrocyte progenitor cell.

Appendix B

Chapter 3

B.1 Sample processing and data generation

B.1.1 Tissue, RNA, and DNA preparation

RNA and DNA were isolated from 50-100 mg of homogenized retina tissue in TRIzol® (Invitrogen, Carlsbad, CA) according to a modified version of the manufacturer's protocol that included additional washing steps. The order of extraction was randomized for all samples. RNA quality and quantity were evaluated using the Bioanalyzer 2100 RNA 6000 Nano assay (Agilent Technologies, Santa Clara, CA). Seven samples with RIN$\leq$5.0 were excluded from the study. DNA was quantified using the QuantiFluor® dsDNA System (Promega, Madison, WI).

RNA library preparation and sequencing

Processing order was randomized before libraries were constructed over two days largely in batches of 24 or 48 with the TruSeq® Stranded mRNA Library Preparation Kit (Illumina, San Diego, CA). The DNA concentration of the sequencing library was determined using the Bioanalyzer DNA 1000 assay (Agilent Technologies, Santa Clara, CA), and a pool of 12 barcoded libraries were layered on a random selection of one of the eight lanes of the Illumina flow cell bridge. Paired-end reads of 125 or 126 base pairs were obtained using the HiSeq 2500 platform (Illumina, San Diego, CA). Sequence data were processed for primary analysis to generate QC values (see Alignment, QC, and quantification below). Samples with a minimum of 10 million mapped reads were retained for downstream analysis.

RNA-Seq QC

Of the 523 samples that were sequenced, twenty-six samples were excluded because of inconsistent or poor subject descriptors as follows: ocular history (1 sample), ambiguous (1) or missing MGS level (5), age at death < 55 years (7), and RIN $<$ 5.0 (12). Six samples were removed after sequencing since < 10 million reads were mapped and/or less than 80% of reads aligned to the reference genome, and 10 samples were eliminated because of skewed gene body coverage over housekeeping genes. Six samples were taken out due to divergence from European (Caucasian) ancestry. Discordant *CFH* and *ARMS2* SNP calls between in-house and Michigan genotyping results were also removed from further analysis (*CFH*: 1 sample; *ARMS2*: 6 samples). Discordance between nominal gender, genetically inferred gender, and gender

inferred from RNA-Seq Principal Component Analysis identified 7 mismatches, and these samples were not used for further analysis. Thus, a total of 70 unique samples were removed, and the entire QC process yielded 453 high-quality samples for gene expression analysis (105 MGS1, 175 MGS2, 112 MGS3, and 61 MGS4).

Alignment, QC, and quantification

Raw RNA-Seq reads were trimmed for Illumina adapters and low quality (SLIDING WINDOW 4:5; LEADING 5; TRAILING 5; MINLEN 25) in Trimmomatic (version 0.36). QC check was performed using FastQC (version 0.11.5). Trimmed reads were aligned to the Ensembl release 85 (GRCh38.p7) human genome using STAR (version 2.5.2a) with per-sample 2-pass mapping and ENCODE standard options. Additional QC metrics were calculated from Trimmomatic, FastQC and STAR using in-house Python and R scripts, including FASTQ and BAM file sizes, total number of reads, number of mapped and unmapped reads, and percentage of mapped reads. RNA-Seq data were also inspected for uniform full-length gene body coverage across housekeeping genes using RSeQC (version 2.6.4). RSEM (version 1.13.1) was used to obtain estimated gene- and transcript expression levels. Normalization was performed using Trimmed Mean of M-values (TMM) in Counts per Million (CPM) using edgeR (version 3.18.1), and then converted into $\log_2$ CPM with an offset of 1. For eQTL analysis, normal quantile transformation was applied instead to $\log_2$(CPM) values. Non-autosomal genes and genes aligning on chromosomal patches/scaffolds were removed from reference transcriptome and eQTL analyses. Expression of cell-type specific markers in the retina did not show any significant changes across MGS

stages, indicating no major loss of cell types during AMD progression (data not shown).

Reference annotation-based assembly

After individual transcriptomes were assembled using the Reference Annotation-based Transcript Assembly method within Cufflinks suite (version 2.21), all assemblies were merged in Cuffmerge and a single, unique set of assembled transcripts was generated using Cuffcompare. Over 91% of transcripts in the reference annotation were captured (196,558 out of 215,929 transcripts), giving a comprehensive general view of the retina transcriptome. This transcript assembly was then processed using the following filters to identify putative lincRNA and protein-coding transcripts: (1) exon count, (2) transcript length, (3) coding potential, (4) functional protein domains, (5) distance to nearest protein-coding gene, and a transcript-level expression threshold at least 1CPM $\geq$ 50% of MGS1 controls.

In order to identify lincRNA, multi-exonic transcripts of at least 200 base pairs were extracted from the transcript assembly. TransDecoder (version 2.0.1) was applied to select for transcripts with a maximum open-reading frame of 75 amino acids lacking coding capacity. CPAT (version 1.2.2) was used as a second independent method to assess coding potential, and only those lincRNA located at least 2 Kb away from the nearest protein-coding gene were retained. In order to determine protein-coding transcripts, all multi-exonic transcripts were extracted from the transcript assembly. TransDecoder was applied to select for transcripts with a minimum

open reading frame of 50 amino acid residues of coding capacity, Pfam-based HMMER (version 3.1.b) (see URLs) was used to retain transcripts with best 1 domain e-value of$\leq$0.05 and at least one known functional domain, and CPAT was implemented to further assess coding potential. The logistic regression model and hexamer table required for CPAT were built using 10,000 coding sequences from the Consensus Coding Sequence Project and 10,000 annotated noncoding sequences from GenCODE (release 25). The model was evaluated with 10-fold cross validation. A two-graph receiver operating characteristic curve was generated to select the optimum coding probability cutoff value (coding $\geq$ 0.3755; noncoding $<$ 0.3755).

B.1.2 Genotyping

DNA from 516 samples, along with replicates as QC for 30 random samples, were genotyped using the UM_HUNT_Biobank v1.0 chip, which is based on the Illumina Infinium CoreExome-24 bead array platform (Illumina, San Diego, CA) with 547,655 markers and an additional 55,939 custom content markers. Genotype analysis was performed with Illumina GenomeStudio (module 1.9.4, algorithm GenTrain 2.0). We also performed TaqMan SNP genotyping for two variants, in *CFH* (Y402H; rs1061170) and *ARMS2* (A69S; rs10490924), using the ABI 7900HT sequence detection system (Applied Biosystems, Foster City, CA). The Y402H variant in *CFH* was assayed using a custom-made probe and the A69S variant in *ARMS2* was analyzed using a commercially available TaqMan probe (C_29934973_20). Briefly, 15-30 ng of DNA was mixed with TaqMan genotyping master mix (Applied Biosystems, Foster City, CA) and TaqMan SNP genotyping assay mix (40X; Applied Biosystems, Foster

City, CA) in a total volume of 15μl. Following PCR, allele discrimination was carried out with the ABI Prism 7900HT genetic detection system (Applied Biosystems, Foster City, CA).

eQTL QC and imputation

Of the genotyped samples, 20 samples were excluded from analysis: missingness>5% in 1 sample, relatedness (2nd degree or higher) in 14 samples, and contradictions in inferred and reported sex in 5 samples. Initial QC at the SNP-level involved (1) removal of SNPs with HWE P value < 1E-06, (2) call rate < 95%, and (3) duplicate and non-autosomal variants. We retained 570,441 variants. Genotypes were imputed with IMPUTE2 based on the 1000 Genomes Project Phase 3 reference panel (October 2014). For our eQTL analysis, QC after imputation excluded: (1) poorly imputed variants (info < 0.3), (2) indels of length>51 bp, (3) imputed variants with HWE < 1E-06, (4) imputed variants with MAF<0.01, and (5) monomorphic variants. In total, 8,924,684 autosomal variants across 406 individuals were retained, and coordinates were then converted from Ensembl GRCh 37.p13 to Ensembl GRCh 38.p7 in order to match the retina RNA-Seq data. Population stratification was examined using Eigenstrat to identify 11 significant principal components; 10 of these were used in the final eQTL model.

B.2 Batch correction

Exclusion criteria for negative control genes in SSVA included: (1) Genes within 100 Kb of linkage disequilibrium of known 34 AMD susceptibility loci identified in the most recent GWAS study for AMD, (2) RetNet (retinal Information Network) genes (see URLs), (3) AMD candidate genes from PubMed literature search over the last five years (see Weighted Gene-correlation Network Analysis in Methods), (4) aging- and gender-associated genes from GTEx analysis, (5) X and Y chromosomal genes, and (6) genes that did not meet the expression-level threshold $\geq$ 1 CPM in $\geq$ 10% of all samples.

B.3 eQTL, TWAS, and eCAVIAR

B.3.1 Enrichment

We examined whether there is a broader relationship between cis-eQTLs and AMD genetic susceptibility beyond what has been observed for known GWAS loci. A Q-Q plot for each of the GWAS datasets was generated by: (1) subsetting to International HapMap Project phase 3 (NCBI build 36, dbSNPb129) variants in the European population with MAF$\geq$0.05, (2) removal of variants in the major histocompatibility complex region, and (3) removal of variants within $\pm$ 1 Mb of the known GWAS signals. We then stratified the variants into multiple (overlapping) categories based on eQTL characteristics: (1) retina-specific eQTLs: eVariants that regulate gene expression only in retina, (2) GTEx-1 eQTLs: eQTLs that regulate gene expression

in at least 1 GTEx tissue (3) GTEx-20 eQTLs: eQTLs that regulate gene expression in at least 20 GTEx tissues, and (4) GTEx-40 eQTLs: eQTLs that regulate gene expression in at least 40 GTEx tissues.

B.3.2 Colocalization

Fine mapping using eCAVIAR was performed in the following manner: (1) for each lead variant in GWAS, a 1Mb window around it was defined as its locus, (2) for all variants within the locus, we identified/defined target genes as genes that are associated at FDR$\leq$0.05 with any of these variants in the eQTL study, and (3) we calculated the colocalization posterior probability (CLPP) for each variant and target gene within the loci. The most relevant target gene was then defined as the gene with the highest CLPP above the threshold of 0.01 within the loci. A maximum of three possible causal variants for each locus was assumed.

B.3.3 TWAS

The TWAS procedure required that we model gene expression with genotype. The gene expressions were modeled using either elastic net, mixed models, or least absolute shrinkage and selection operator (LASSO). The LASSO lambda parameter was calculated using the heritability; genes for which the heritability could not be calculated used the average heritability across genes instead. Of the 18,053 genes expressed in the retinal samples, 17,345 were present in the TWAS analysis. Genes not analyzed in TWAS were located on either sex chromosomes, the mitochondrial chromosome, on scaffolds, or did not have SNPs within 1 Mb of the merged GWAS-eQTL

SNP set. The mean cross-validated model fit was 0.07, and the mean heritability of the 14,353 genes for which it could be calculated was 0.127. As expected, the higher the heritability, the better the cross-validated model fit. LASSO was the best fit for approximately half of the genes, and elastic net accounted for another quarter; genes for which the mixed model provided the best fit had models that captured less variation in expression than other genes.

The TWAS statistics does not take into account LD between genes, so we performed summary-level equivalent conditional tests for each chromosome for genes that were both significant at an FDR of 0.05 and had a genetic expression model R^2 >than 0.01. Genes were added in a stepwise manner into the model, from lowest marginal p-value to highest, until no gene remained significant. The model prior to this saturation was used as the final conditional model; no provision was made to prevent over-fitting. Of the 61 genes tested, 47 remained nominally significant at $\alpha = 0.05$; of these, 39 remained significantly associated after Bonferroni correction for multiple testing (using all 61 genes considered for the test, not just ones included in the models). A permutation test (described in Methods) was also performed; seven genes were significant after Bonferroni correction and had a gene model R^2 >0.01, and three of these were outside of the GWAS loci: PARP12, MTMR10, and SH3BGR.

We explored the tissue specificity of these results, at least in part, using GTEx data v6. We downloaded the pre-computed TWAS weights derived from the data of 39 GTEx tissues (excluding cell lines and biological replicate tissues, such as frontal cortex and cerebellar hemisphere) from the TWAS website (http://gusevlab.org/projects/fusion/)

and performed the procedure for the GTEx weights with the same set of AMD GWAS summary statistics that was used with retina. The complete results of the TWAS analysis - gene model attributes, marginal association statistics, conditional and permutation test results, and GTEx marginal associations for the retinal candidates with FDR<0.05 - are provided in Supplementary Data S5. Please note that relatively few genes had weights available in most GTEx tissues.

B.3.4 Evaluation of AMD GWAS lead variants for eQTL evidence in non-retina tissues

Of the 52 lead variants from AMD-GWAS, 41 were analyzed in our study. Those not found were either not in the reference dataset used for imputation (6 variants) or did not pass our MAF threshold (5 variants, MAF threshold; 0.01). Matrix eQTL was then used to obtain the marginal associations using the same cis criteria, which were then corrected for multiple testing using the Bonferroni approach at the type I error rate of 0.05.

We compared our findings to that of Strunz et al. which includes eQTLs from liver samples of 588 individuals and GTEx (v7). For the Strunz et al. comparison, we used 31 SNPs with MAF$\geq$ 0.05 that were common to both studies. For each variant, eQTL analysis was performed for all genes that are present within a 1Mb window and expressed in the two tissues (Supplementary Data 3). We also tested 37 AMD-associated variants (with MAF$\geq$ 0.01) that were analyzed in the retina and detected in at least one GTEx (v7) tissue. For each SNP-gene combination, we list all the GTEx tissues that had p-values less than or equal to that of retina

(Supplementary Data 3), or if no GTEx tissue had p-value lower than the retina, we listed all tissues with their respective p-values.

B.4 Gene expression analysis

B.4.1 GSEA

We focused on gene sets that passed a significance threshold of FDR q-value ≤ 0.25 and on key genes that appeared in at least 25% of gene sets in common functional categories using Leading Edge Analysis (Supplementary Data S6). Comparison of early AMD to controls identified 38 significantly enriched gene sets, all upregulated and generally relating to cell killing (3), metabolism (12), and the immune system (15). The largest of these categories involved immune system processes (13) with an average normalized enrichment score (avg. NES) of 2.4 and 80 key genes. Comparison of intermediate AMD to controls identified 6 upregulated and 60 downregulated significantly enriched gene sets comprising metabolism (2), cell killing (2), and cellular component organization (3). Comparison of advanced AMD to controls identified 44 upregulated and 15 downregulated significantly enriched gene sets including those relating to metabolism (21), cell component organization (9), immune system (6), and stress response (4). Additionally, we identified downregulated gene sets that were predominant and largely exclusive to intermediate AMD and associated with synapses in cell communication (14, avg. NES = -2.2), nervous system development (9, avg. NES = -2.4), biological regulation (4, avg. NES = -2.3), and establishment/maintenance of cell polarity (3, avg. NES = -2.4) (Supplementary Table S6).

B.4.2 Comparison of transcriptomes across retina and GTEx tissues

The bioinformatics pipeline used to analyze RNA-Seq data in this study mainly differed from that of GTEx v7 in gene quantification methods and gene annotation version. To understand the consequences of using different pipelines and to ensure appropriate tissue comparisons between studies, multidimensional scaling plots and hierarchical clustering dendrograms were generated based on normalized gene expression levels from the different pipelines. Statistical methods used to generate the multi-dimensional scaling (MDS) plot itself were obtained from GTEx. Three comparisons were made based on the following data sets: (1) Raw GTEx v7 data processed through our pipeline, (2) Raw GTEx v7 and retina data processed through our pipeline, and (3) GTEx v7 gene-level TPM count data provided on the GTEx online portal.

Raw GTEx v7 data were processed through our pipeline as previously mentioned in 1.2.2 Alignment, QC, and quantification. In addition, we used similar methods that GTEx had applied to detect samples outliers. PCA-based outlier detection was performed in the first two principal components by using Mahalanobis distance to center the data. Outliers were identified using a threshold of three standard deviations.

B.5 Supplementary Figures and Table

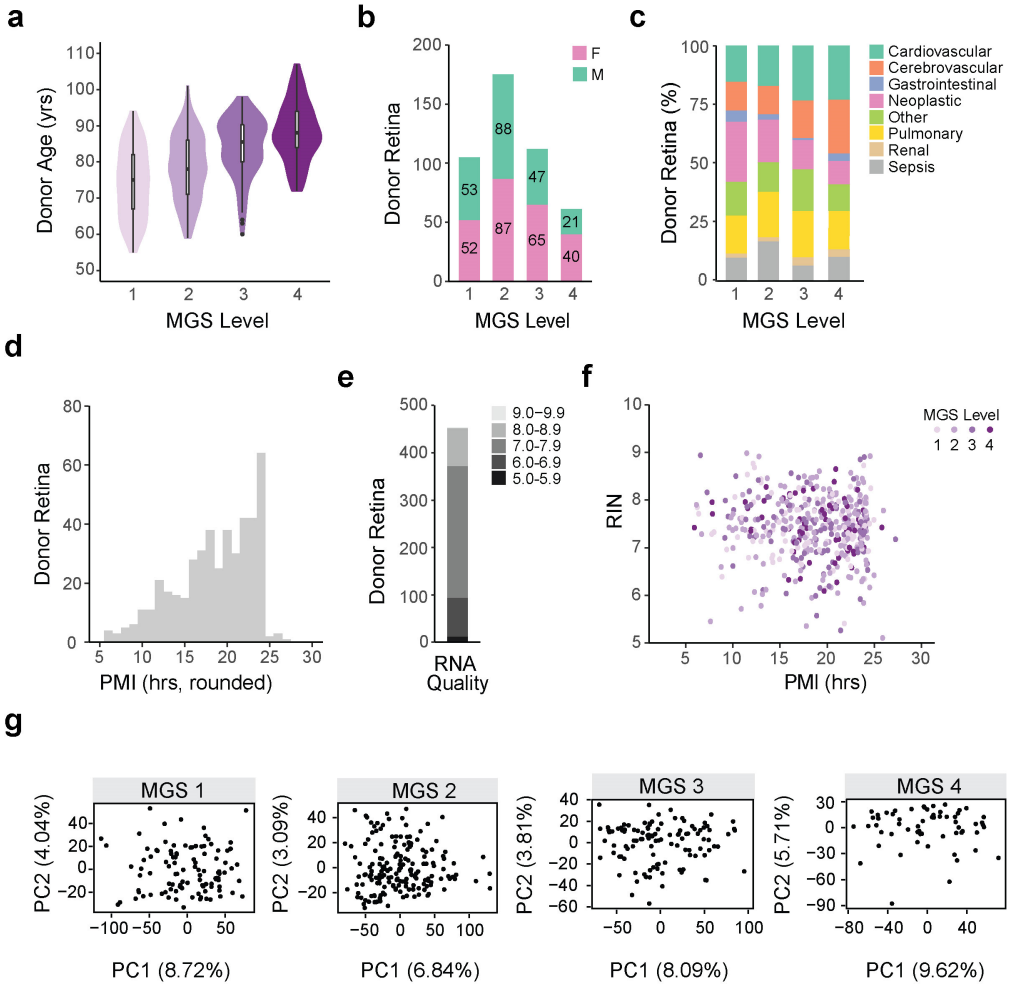


Fig. B.1. Characteristics of retina donor samples used in this study. **a**, Violin plots showing the age distribution, in years, of donors across the four MGS stages. The boxplot within each violin plot depicts the median, and the lower and upper hinges correspond to the first and third quartiles, respectively. Outlying data are represented by individual points that extend beyond 1.5 × interquartile range below the first quartile or above the third quartile. The mean age of donors was 80 years (range 55-107), and the mean donor age increased with AMD severity: 74 years (range 55-94) in MGS1, 78 years (59-101) in MGS2, 84 years (60-98) in MGS3, and 88 years (range 72-107) in MGS4. **b**, Distribution of gender across the four MGS stages. Gender was distributed almost evenly in MGS1 to MGS3, with almost twice as many females as males in MGS4. **c**, The cause of death across the four MGS stages. Donors within each MGS stage were grouped into 8 categories based on the reported cause of death to determine that causes of death were not conflated with donor age or MGS stage. **d**, Distribution of post-mortem interval (PMI), in hours. PMI was defined as the mean time lapse from death to enucleation and tissue cryopreservation. Mean PMI was 18.66 hours. **e**, Quality of RNA, as defined by the RNA Integrity Number (RIN), used for RNA-Seq. Mean RIN was 7.42±0.6 (5.1-9). **f**, Scatterplot of RNA integrity (RIN) versus post-mortem interval (PMI). **g**, PCA plots of donors within each MGS level based on normalized gene expression levels.

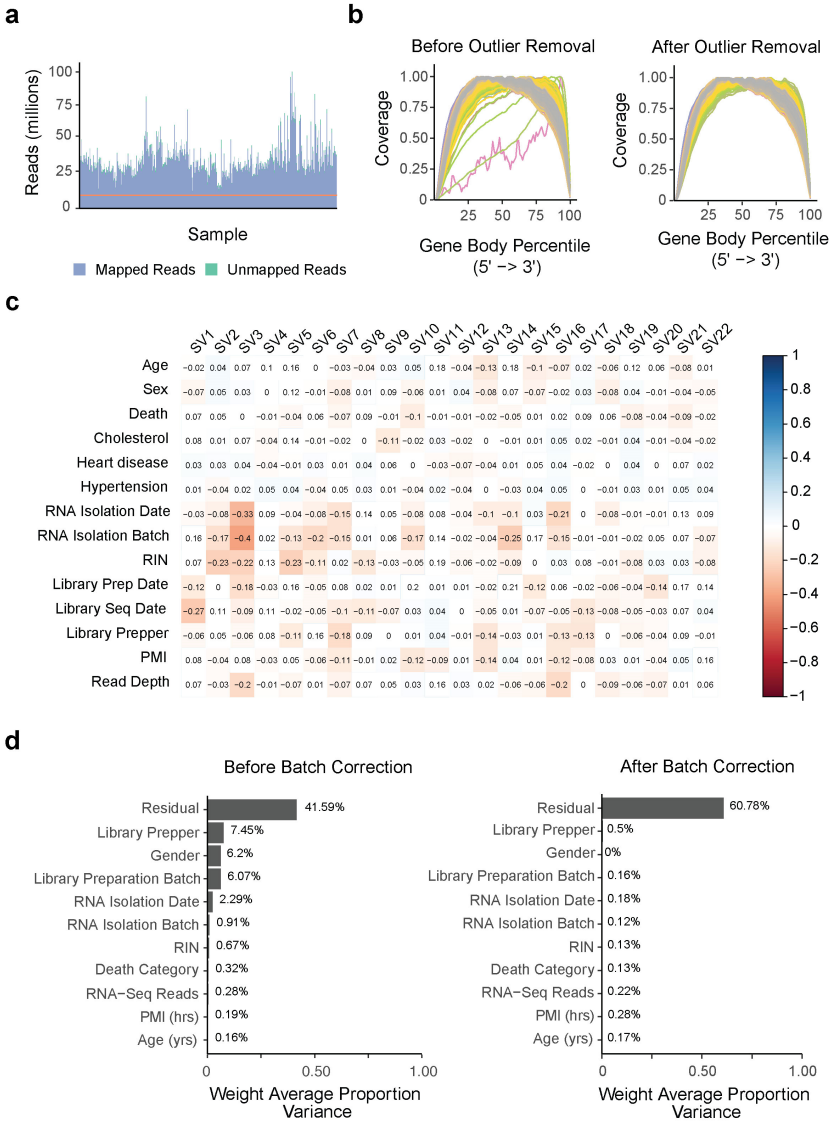


Fig. B.2. RNA-Seq QC metrics. **a**, Number of RNA-Seq reads that mapped to the human reference genome Ensembl 38.85. The red horizontal line denotes 10 million reads. **b**, Normalized mean per-base 5' to 3' gene body coverage of housekeeping genes. Left: before outlier removal. Right: after outlier removal. **c**, Correlation between 22 significant surrogate variables identified in SSVA and possible documented sources of variation. A p-value of 0.05 was used as the significance threshold. Correlation coefficients are labeled in black and color-coded such that positive correlations are displayed in blue and negative correlations in red. Color intensity is proportional to the correlation coefficients. RIN: RNA Integrity Number; PMI: post-mortem interval. **d**, Principal variance component analysis (PVCA) of the retina gene expression data set. Residual represents the remaining variance in the data set not attributed to the specified batch and biological variables. Left: before batch correction. Right: after batch correction. RIN: RNA Integrity Number; PMI: post-mortem interval.

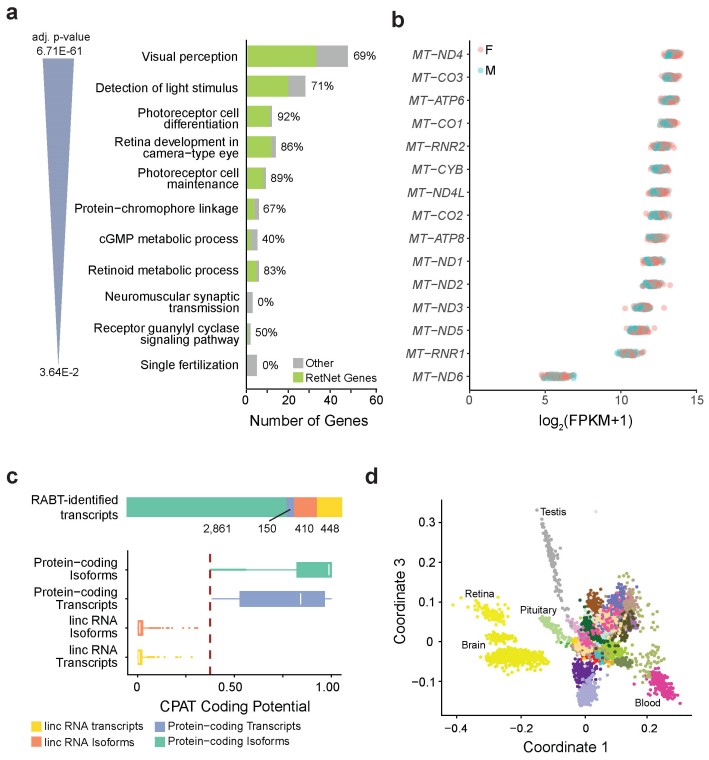


Fig. B.3. Reference transcriptome of the human retina. **a**, Gene Ontology (GO) Biological Process pathway enrichment analysis of high abundance genes ($\geq$100 FPKM) in the retina. The bars represent the number of genes identified in each pathway, highlighting in green the number of inherited retinal disease-causing genes in the RetNet database of ocular diseases (percentage indicated to the right of bar). Redundancy of enriched GO terms was removed using a similarity cutoff of 0.40. A Benjamini-Hochberg adjusted p-value$\leq$0.05 was used as the significance threshold. **b**, Scatter plot of mitochondrial gene expression based on $\log_2$(FPKM+1) values among males and females. **c**, Novel transcript discovery using reference annotation-based transcript assembly. Top: Number of putative novel protein-coding and lincRNA isoforms and transcripts. Bottom: Coding Potential Assessment Tool (CPAT) coding probability score of putative novel protein-coding and lincRNA isoforms and transcripts. The dotted red vertical line denotes the calculated coding probability cutoff of 0.3755. We discovered a total of 410 and 2,861 lincRNA and protein-coding isoforms, respectively, and a total of 150 and 448 lincRNA and protein-coding transcripts, respectively. **d**, Multidimensional scaling plot of samples across tissues based on normalized gene expression levels.

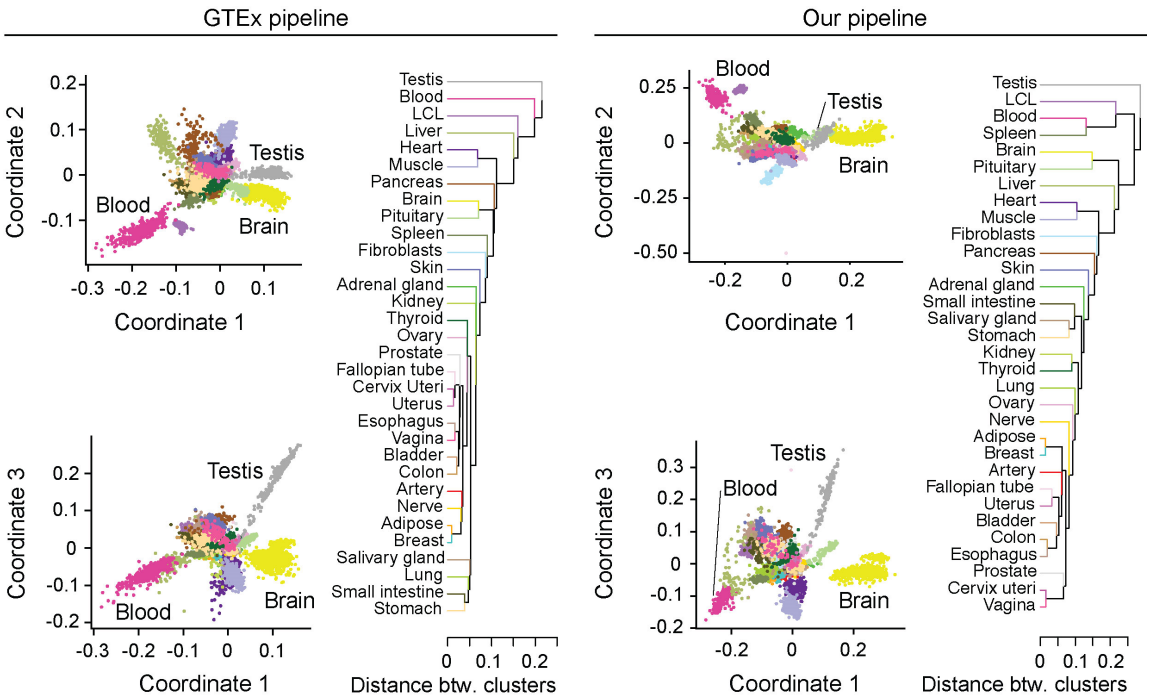


Fig. B.4. Comparison of RNA-Seq analysis pipelines using GTEx data without retina. Multidimensional scaling plots and hierarchical clustering dendrograms of samples across tissues based on normalized gene expression levels. Left: based on our bioinformatics pipeline. Right: based on GTEx v7 gene-level TPM count data. These comparisons suggest that the relationship between tissues was not affected by the analysis pipeline.

Our RNA-seq analysis pipeline was based on the most recent literature recommendations for RNA-Seq analysis (as described in Methods) and mainly differed from that of GTEx in gene quantification methods and in gene annotation version. We therefore downloaded the raw GTEx data and processed these through our bioinformatics pipeline to generate the MDS plot. Statistical methods used to generate the MDS plot itself were obtained from GTEx. In addition, we explored whether similar findings could be obtained using a different analysis pipeline. We also plotted MDS plots from expression data provided on the GTEx online portal. MDS plots and hierarchical clustering dendrograms generated from different pipelines were comparable.

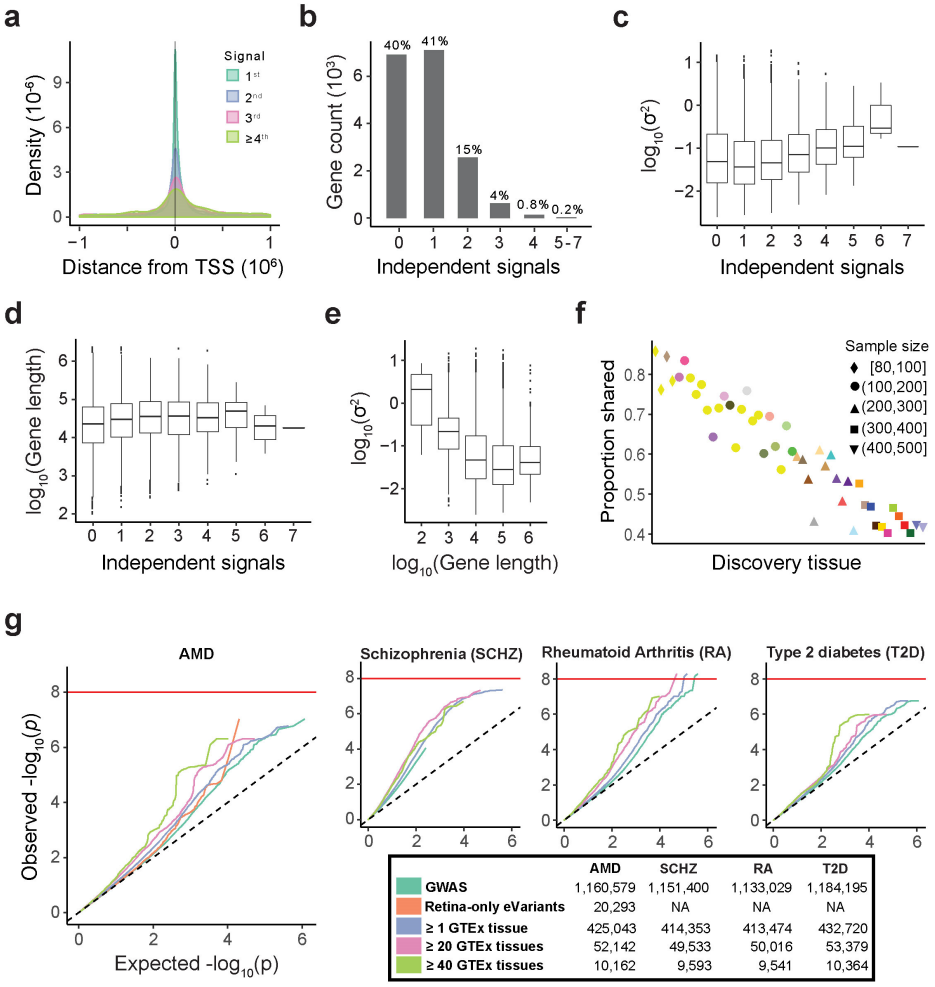


Fig. B.5. cis-eQTL analysis. **a**, The relationship between the strength of each cis-eQTL's association and the distance of its eVariant from its eGene's transcription start site (TSS). **b**, The distribution of cis-independent signals for each autosomal gene. Thus approximately 60% of genes in the retina were found to be under genetic control with the majority of the genes having one independent signal (41%). **c**, Distribution of the amount of variability left unexplained in gene expression levels after correction for other covariates used in the model stratified by the number of independent signals found per gene. **d**, Distribution of gene length stratified by the number of independent signals found per gene. **e**, Distribution of the amount of variability left unexplained in gene expression levels after correction for other covariates used in the model ordered by gene length. **f**, Proportion of cis-eQTLs discovered in GTEx that were replicated in the retina (y-axis), ordered by sample size in discovery tissue (x-axis). The color and shape of the points represent the sample size of the replication tissue. **g**, Q-Q plot indicating the relationship between the observed -$\log_{10}$ p-values for each stratum relative to its expected null distribution. Each stratum, except for the GWAS one, classifies the eVariants by how many tissues they regulate at least one gene in. This analysis is shown for AMD, schizophrenia, rheumatoid arthritis, and Type 2 diabetes.

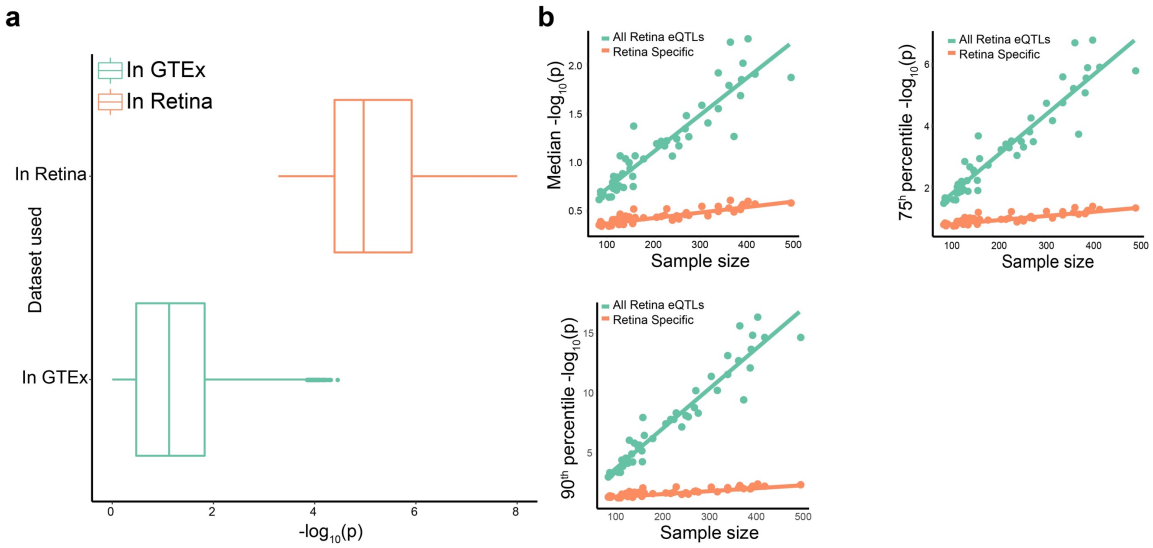


Fig. B.6. Comparison of retina-specific eQTLs across GTEx. **a**, Boxplots showing minimum p-values across GTEx tissues for eQTLs detected only in the retina, after correcting for the number of tissues eQTLs were tested in. As a comparison, distribution of p-values in the retina analysis for the same eQTLs are also shown. The distribution of p-values between retina and other tissues is expected given that these SNPs, by definition, are significant eQTLs in retina, but not in other tissues. **b**, Median, 75th, and 90th percentile of -$\log_{10}$(p-values) of retina-specific cis-eQTLs in different non-retina tissues against their respective sample sizes. These plots were generated to explore whether SNPs that were not detected as significant eQTLs in non-retina tissues using the stringent p-value threshold could still reveal some enrichment towards lower p-values than what is expected by chance. We also compared this trend for all eQTLs detected, regardless of whether they were retina-specific or not. A weak trend towards lower p-values in tissues with large sample sizes for retina-specific eQTLs was observed. However, this trend was much weaker compared to that observed for all eQTLs. It appears that retina-specific eQTLs have stronger effects in the retina though possibility of weak effects of these eQTLs in other tissues cannot be ruled out.

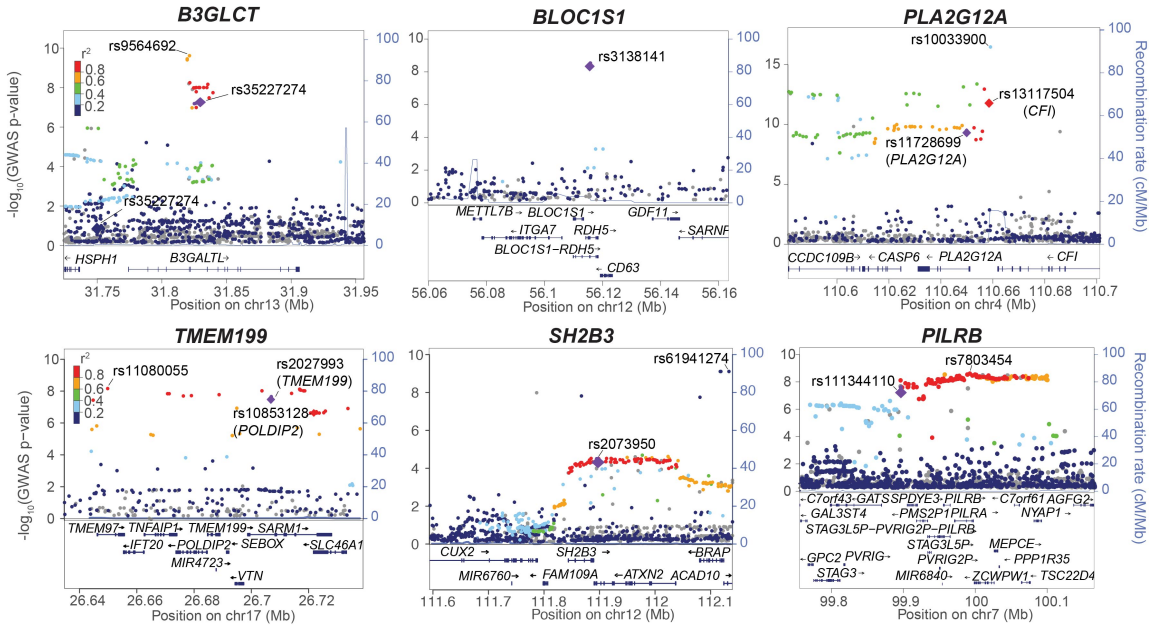


Fig. B.7. Manhattan plots at known AMD loci. LocusZoom [30] -generated Manhattan plot of GWAS regions encompassing the candidates that fell within known AMD loci and were shown to be associated through multiple methods of analysis, as specified by Table 1. The top variants for the independent eQTL signals determined by the conditional analysis are displayed as diamonds and labeled. The SNP with the strongest GWAS signal in the region is also identified in each plot. Coloration of the points is determined by strength of linkage disequilibrium (LD) with respect to the top variant of the strongest eQTL signal. If LD information provided to LocusZoom was absent for that SNP, one of its proxies according to LDLink [31] (R^2 >0.99) was used. Recombination rate is shown as a blue line.

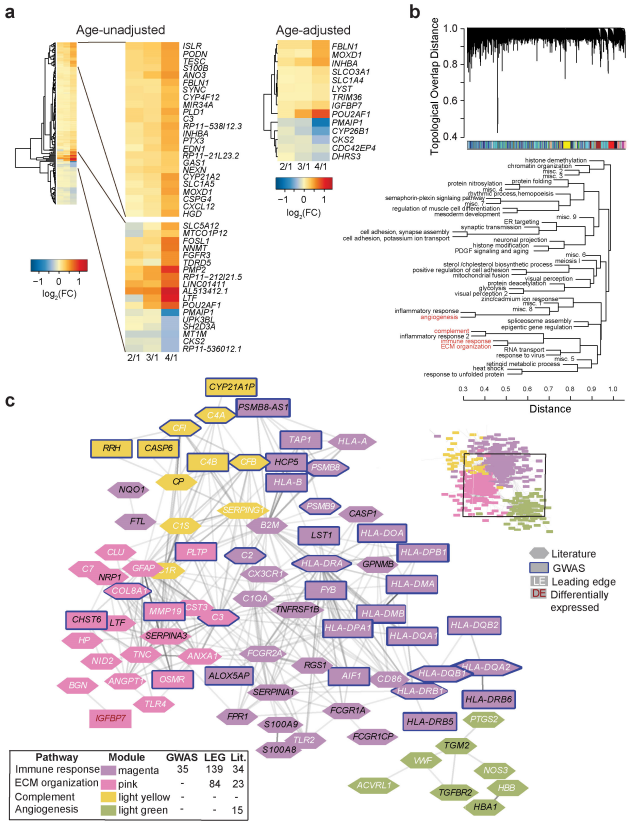


Fig. B.8. Differential expression and WGCNA analysis. **a**, Heatmap showing the expression pattern of differentially expressed genes by comparing advanced AMD to controls with and without adjusting for age at the significance threshold at FDR≤0.20 **b**, We identified 47 modules, each containing between 16 and 4,847 genes. Top: Dendrogram of genes with topological overlap used as distance (shown on y-axis). The color bar below indicates which module the genes belong to. Bottom: Hierarchical clustering of module expression eigenvalues (eigengenes). The modules involved in complement (yellow), angiogenesis (light green), immune activation (magenta), and extracellular matrix (pink) are highlighted in red. These modules were adjacent to each other according to eigenvalue-based hierarchical clustering. **c**, Two of these modules were particularly interesting as they were enriched for literature (pink FDR= 2.21×10^{-3}; magenta FDR= 1.37^{-9}) and leading edge (pink FDR = 1.10×10^{-3}; magenta FDR= 1.33×10^{-26}) candidate genes. Additionally, the magenta module was enriched for genes from the GWAS loci (FDR= 2.38×10^{-4}). The pink module also contained three DE- (FBLN1, MOXD1, IGFBP7) and two AMD-associated genes (COL8A1 and MMP19). GO analysis of the magenta and pink module highlighted extracellular matrix organization and immune response pathways, respectively, which were previously implicated in AMD pathology. These modules interacted closely with two other modules; the light green (also enriched for literature genes, FDR= 8.30×10^{-3}) and light yellow, which were enriched for angiogenesis and complement GO terms, respectively. We show only genes that fall in either literature, GWAS, or differentially expressed groups and are strongly correlated with another such candidates (adjacency> 0.05).

Table B.1. Summary of eQTL, eCAVIAR and TWAS analyses for prioritizing variants and target genes across AMD-GWAS loci.

AMD Locus	Lead GWAS SNP	Chr:Position	GWAS *P*	eQTL *P*	Target gene(s)	Percent variability explained	Significant TWAS genes in the locus (FDR < 0.05)
B3GALTL	rs9564692	13:31821240	3.31×10^{-10}	2.36×10^{-11}*	B3GLCT†	10.47	B3GLCT (1.37×10^{-4})
RDH5/CD63	rs3138141	12:56115778	4.3×10^{-9}	5.69×10^{-19}*	BLOC1S1†, RP11-644F5.10	17.80	BLOC1S1 (6.36×10^{-6}), RP11-644F5.10 (2.89×10^{-6})
SLC16A8	rs8135665	22:38476276	5.53×10^{-11}	1.56×10^{-3}	CTA-228A9.3†	2.45	CTA-228A9.3 (1.26×10^{-5})
ACAD10	rs61941274	12:112132610	1.07×10^{-9}	8.95×10^{-2}	SH2B3†	0.71	SH2B3 (2.16×10^{-2})
PILRB/PILRA	rs7803454	7:99991548	4.76×10^{-9}	3.57×10^{-77}*	PILRB, STAG3L5P, PILRA, ZCWPW1, TSC22D4	57.51	MEPCE (5.83×10^{-6}), PMS2P1 (1.11×10^{-5}), STAG3L5P-PVRIG2P-PILRB (1.88×10^{-5}), PILRB (1.88×10^{-5})
TMEM97/VTN	rs11080055	17:26649724	1.04×10^{-8}	8.37×10^{-19}*	POLDIP2, SLC13A2**, TMEM199†	17.65	TMEM199 (2.37×10^{-5}), POLDIP2 (8.27×10^{-5})
CFI	rs10033900	4:110659067	5.35×10^{-17}	3.98×10^{-7}*	PLA2G12A	6.17	CFI (3.31×10^{-10}), PLA2G12A (4.53×10^{-10})
KMT2E/SRPK2	rs1142	7:104756326	1.35×10^{-9}	6.49×10^{-6}*	CTB-152G17.6**	4.91	
NPLOC4/TSPAN10	rs6565597	17:79526821	1.45×10^{-11}	1.91×10^{-5}*	ARL16	4.43	ANAPC11‡ (4.03×10^{-3})
C2/CFB/SKIV2L	rs114254831	6:32155581	9.4×10^{-12}	4.70×10^{-6}*	HLA-DQB1	5.06	SKIV2L (1.78×10^{-31})
APOE	rs429358	19:45411941	2.39×10^{-42}	2.85×10^{-3}	CTB-129P6.7, TOMM40†	2.18	
APOE	rs73036519	19:45748362	3.14×10^{-7}	3.80×10^{-2}	ZNF180, TOMM40†	1.06	
C2/CFB/SKIV2L	rs116503776	6:31930462	1.17×10^{-103}	3.71×10^{-4}	DXO	3.09	SKIV2L (1.78×10^{-37})
CETP	rs5817082	16:56997349	3.56×10^{-19}	1.18×10^{-3}	NLRC5	2.57	HERPUD1 (9.66×10^{-5})
CETP	rs17231506	16:56994528	2.18×10^{-18}	6.56×10^{-3}	HERPUD1	1.81	HERPUD1 (9.66×10^{-5})
COL8A1	rs55975637	3:99419853	1.30×10^{-8}	1.32×10^{-2}	NIT2	1.51	TOMM70 (2.55×10^{-2})
CFH	rs10922109	1:196704632	9.6×10^{-618}	7.44×10^{-3}	KCNT2	1.76	KCNT2 (1.04×10^{-20})
CFH	rs570618	1:196657064	2.0×10^{-590}	1.42×10^{-2}	CFH	1.48	KCNT2 (1.04×10^{-20})
CFH	rs187328863	1:196380158	1.06×10^{-68}	2.63×10^{-2}	ZBTB41	1.22	KCNT2 (1.04×10^{-20})
CFH	rs61818925	1:196815450	6.03×10^{-165}	3.21×10^{-1}	ZBTB41	0.24	KCNT2 (1.04×10^{-20})
CNN2	rs67538026	19:1031438	2.58×10^{-8}	9.21×10^{-11}*	TMEM259	9.87	
RAD51B	rs61985136	14:68769199	1.56×10^{-10}	2.15×10^{-2}	PIGH	1.30	TMEM229B (2.64×10^{-2})
RAD51B	rs2842339	14:68986999	1.36×10^{-6}	5.60×10^{-2}	ZFYVE26	0.90	TMEM229B (2.64×10^{-2})
MMP9***	NA	NA	NA	NA	NA	NA	PLTP (3.3×10^{-2})

*eQTL is significant after correction for multiple testing. **Retina-specific. ***Lead SNP not present in the dataset, and suitable proxy SNPs are not available. †Gene is target of causal variant identified by eCAVIAR. ‡Low TWAS model fit ($R^2 < 0.01$).

Appendix C

Chapter 4

C.1 Definitions/Assumptions

1. We assume the polygenic model for trans-eQTLs. This is written as

$$Y_g = X\beta_g^{(J)} + \epsilon \tag{C.1}$$

representing the model for the gth gene. Where Y_g is a N $\times$ 1 vector of gene expression levels and X is a N $\times$ K_g matrix of genotypes for K_g SNPs across N subjects. Under this approach, we assume that the distribution of the true effect sizes, based on the joint model above, across the K_g SNPs for a given gene, g, are i.i.d according to the following distribution

$$\beta_{kg}^{(J)} | \sigma_g^2, \pi_g \sim \pi_g \mathcal{N}(0, \sigma_g^2) + (1 - \pi_g)\delta_0 \tag{C.2}$$

Where both σ_g^2 and π_g vary over genes and δ_0 is the Dirac delta function indicating a fraction, $1-\pi_g$, of the SNPs have no association with the expression level of gene g.

2. The usual marginal model for trans-eQTLs for the kth SNP is written as

$$Y_g = X_k \beta_{kg}^{(M)} + \epsilon$$

representing the model for the kth SNP. The true marginal effect size of the k SNPs on the gth gene is related to its joint effect on the same gene accounting for the other snps in the model by the following relationship (for clarity, we omit the notation g, and note that each snp effect is gene specific so that the form below depends on SNPs fitted for the same gene)

$$\beta_k^{(M)} = \sum_{p=1}^{P} \beta_p^{(J)} \rho_{kp} \tag{C.3}$$

where ρ_{kp} is the Pearson correlation coefficient between SNP k and p.

3. Conditional on the true marginal effect size the (marginal) OLS estimate follows a normal distribution, shown below

$$\hat{\beta}_{kg}^{(M)} | \beta_{kg}^{(M)} \sim \mathcal{N}(\beta_{kg}^{(M)}, a + s_{kg}^2)$$

where the factor "a" is introduced to account for possible systematic bias in variance estimates due to effects such as population stratification or cryptic

relatedness with respect to gene g.

4. We place the following priors on π_g and σ_g^2.

$$\pi_g \sim \text{Beta}(\alpha_0, \beta_0) \text{ and } \sigma_g^2|\pi_g \sim \text{Inverse Gamma}(a_0, b_0)$$

set $a_0 = \nu_0/2$ and $b_0 = \nu_0\sigma_0^2/2$.

5. In general if $X|\sigma^2 \sim \mathcal{N}(0, \kappa\sigma^2)$ and $\sigma^2 \sim$ Inverse Gamma (a_0, b_0), where $a = \nu_0/2$ and $b_0 = \nu_0\sigma_0^2/2$, then X has the Student-t marginal distribution shown below

$$\frac{\Gamma\left(\frac{\nu_0+1}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)\sqrt{\kappa\nu_0\sigma_0^2\pi}}\left(1+\frac{1}{\kappa\nu_0}\left[\frac{x}{\sigma_0}\right]^2\right)^{-\left(\frac{\nu_0+1}{2}\right)}$$

6. Let $Y|\mu \sim \mathcal{N}(\mu, \sigma^2)$ and μ be t-distributed with location parameter zero and scale parameter k with ν degrees of freedom, then the distribution of Y marginalized over μ is given as

$$f_Y(y) = \frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu k^2}}\left(\frac{\nu k^2}{2\sigma^2}\right)\left(\frac{\nu+1}{2}\right)\int_0^\infty \frac{t^{[(\nu+1)/2]-1}}{\sqrt{1+t}}\exp\left(\frac{1}{2\sigma^2}\left[\frac{y^2}{1+t}-t\nu k^2\right]\right)\,dt$$

and can be simplified to give

$$f_Y(y) = \frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi\sigma^2}}\left(\frac{\nu k^2}{2\sigma^2}\right)^{\nu/2}\Gamma\left(\frac{\nu+1}{2}\right)\sum_{i=0}^{\infty}\frac{1}{i!}\left(\frac{y^2}{2\sigma^2}\right)^i \cdot U\left(\frac{\nu+1}{2}, \frac{\nu}{2}+1-i, \frac{\nu k^2}{2\sigma^2}\right)$$

Where U(a,b,z) is Tricomi's (confluent hypergeometric) function and is defined

as

$$U(a,b,z) = \frac{1}{\Gamma(a)} \int_0^\infty \exp(-zt) t^{a-1} (1+t)^{b-a-1} \, dt, \text{ (Re a} > 0 \text{)}$$

C.2 Results

1. Let $\mathcal{S}_{kg}$ represent the set of SNPs in the reference panel that may be "tagged" by the kth SNP with respect to gene g[1]. Define $N_{kg} = |\mathcal{S}_{kg}|$ as the number of SNPs in $\mathcal{S}_{kg}$, and $N_{kg}^{(1)}$ as the number of SNPs in $\mathcal{S}_{kg}$ with non-zero effects on gene g based on the joint model. Furthermore, let $\ell_{kg} = \sum_{p \in \mathcal{S}_{kg}} \rho_{kp}^2$. I.e., the linkage disequilibrium (LD) score for SNP k with respect to the set of SNPs in $\mathcal{S}_{kg}$. We have from [105] that under the assumption of independence between LD patterns and probability of a SNP having a non-zero effect on gene, the marginal distribution for $\beta_{kg}^{(M)}$, for SNP k and gene g, is given as

$$\beta_{kg}^{(M)} | \pi_g, \sigma_g^2 \sim \sum_{n_{kg}^{(1)}} f_{N_{kg}^{(1)} | \pi_g, \sigma_g^2}(n_{kg}^{(1)}) \mathcal{N}\left(0, \sum_{h=0}^{1} \frac{n_{kg}^{(h)}}{n_{kg}} \sigma_{h,g}^2 \ell_{kg}\right) \tag{C.4}$$

where $n_{kg}^{(0)} + n_{kg}^{(1)} = n_{kg}$, the observed N_{kg}

$$f_{N_{kg}^{(1)} | \pi_g, \sigma_g^2}(n_{kg}^{(1)}) = \frac{n_{kg}!}{n_{kg}^{(1)}! n_{kg}^{(0)}!} \pi_g^{n_{kg}^{(1)}} (1-\pi_g)^{n_{kg}^{(0)}}, \quad n_{kg}^{(1)} = 0, \ldots, n_{kg}$$

[1]Recall that trans-eVariants have to be a certain distance away from the gene. Hence, for some SNPs not all variants within its neighborhood will be in $\mathcal{S}_{kg}$

Furthermore, from [105] we have that

$$\hat{\beta}_{kg}^{(M)}|\pi_g, \sigma_g^2 \sim \sum_{n_{kg}^{(1)}} f_{N_{kg}^{(1)}|\pi_g,\sigma_g^2}(n_{kg}^{(1)})\mathcal{N}\left(0, \sum_{h=0}^{1} \frac{n_{kg}^{(h)}}{n_{kg}}\sigma_{h,g}^2\ell_{kg} + a + s_{kg}^2\right) \tag{C.5}$$

The goal is to obtain the unconditional distribution of $\hat{\beta}_{kg}^{(M)}$ with respect to σ_g^2 and π_g. Hence (C.5) can be rewritten as

$$\begin{aligned}\hat{\beta}_{kg}^{(M)}|\pi_g, \sigma_g^2 \sim\ & f_{N_{kg}^{(1)}|\pi_g,\sigma_g^2}(0)\mathcal{N}\left(0, a + s_{kg}^2\right) + \\ & \sum_{n_{kg}^{(1)}=1} f_{N_{kg}^{(1)}|\pi_g,\sigma_g^2}(n_{kg}^{(1)})\mathcal{N}\left(0, \sum_{h=0}^{1} \frac{n_{kg}^{(h)}}{n_{kg}}\sigma_{h,g}^2\ell_{kg} + a + s_{kg}^2\right)\end{aligned} \tag{C.6}$$

Case 1; $n_{kg}^{(1)} \geq 1$

We have

$$\int_{\pi_g}\int_{\sigma_g^2}\left(\hat{\beta}_{kg}^{(M)}|\pi_g,\sigma_g^2\right)f_{\sigma_g^2,\pi_g}(\sigma_g^2,\pi_g) = \int_{\pi_g}\int_{\sigma_g^2}\left[\sum_{n_{kg}^{(1)}=1} f_{N_{kg}^{(1)}|\pi_g}(n_{kg}^{(1)}) \times \right.$$

$$\left.\mathcal{N}\left(0,\sum_{h=0}^{1}\frac{n_{kg}^{(h)}}{n_{kg}}\sigma_{h,g}^2\ell_{kg}+a+s_{kg}^2\right)\right] f_{\sigma_g^2,\pi_g}(\sigma_g^2,\pi_g)$$

$$= \int_{\pi_g}\int_{\sigma_g^2}\left[\sum_{n_{kg}^{(1)}=1} f_{N_{kg}^{(1)}|\pi_g}(n_{kg}^{(1)})\times\right.$$

$$\left.\int_{\beta_{kg}^{(M)}} f_{\hat{\beta}_{kg}^{(M)}|\beta_{kg}^{(M)}}\left(\hat{\beta}_{kg}^{(M)}\right) f_{\beta_{kg}^{(M)}|\sigma_g^2}\left(\beta_{kg}^{(M)}\right)\right] f_{\sigma_g^2,\pi_g}(\sigma_g^2,\pi_g)$$

$$= \sum_{n_{kg}^{(1)}=1}\int_{\beta_{kg}^{(M)}}\int_{\pi_g}\int_{\sigma_g^2} f_{N_{kg}^{(1)}|\pi_g}(n_{kg}^{(1)}) f_{\hat{\beta}_{kg}^{(M)}|\beta_{kg}^{(M)}}\left(\hat{\beta}_{kg}^{(M)}\right) \times$$

$$f_{\beta_{kg}^{(M)}|\sigma_g^2}\left(\beta_{kg}^{(M)}\right) f_{\sigma_g^2,\pi_g}(\sigma_g^2,\pi_g)$$

The last line is as a result of the Fubini-Tonelli theorem. Therefore we proceed as follows.

2. Using eqn C.4 and definition (5) we have the unconditional distribution for

$\beta_{kg}^{(M)}$, with respect to σ_g^2, as

$$\beta_{kg}^{(M)}|\pi_g \sim \sum_{n_{kg}^{(1)}} f_{N_{kg}^{(1)}|\pi_g}(n_{kg}^{(1)})\mathcal{D}_{\tau_{kg}}(\sigma_0, \nu_0) \tag{C.7}$$

where

$$\mathcal{D}_{\tau_{kg}}(\sigma_0, \nu_0) = \frac{\Gamma\left(\frac{\nu_0+1}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)\sqrt{\nu_0\tau_{kg}\sigma_0^2\pi}}\left(1+\frac{1}{\nu_0}\left[\frac{\beta_{kg}^{(M)}}{\sigma_0\sqrt{\tau_{kg}}}\right]^2\right)^{-\left(\frac{\nu_0+1}{2}\right)}$$

and $\tau_{kg} = \dfrac{n_{kg}^{(1)}}{n_{kg}}\ell_{kg}$. I.e., $\mathcal{D}_{\tau_{kg}}(\sigma_0, \nu_0)$ is a generalized t distribution with mean 0 for $\nu_0 > 1$, and variance $\tau_{kg}\sigma_0^2\dfrac{\nu_0}{\nu_0-2}$ for $\nu_0 > 2$.

3. From eqn (C.7) the unconditional distribution with respect to π_g is then

$$\beta_{kg}^{(M)} \sim \sum_{n_{kg}^{(1)}} f^*_{N_{kg}^{(1)}}(n_{kg}^{(1)})\mathcal{D}_{\tau_{kg}}(\sigma_0, \nu_0) \tag{C.8}$$

where

$$\begin{aligned} f^*_{N_{kg}^{(1)}}(n_{kg}^{(1)}) &= \binom{n_{kg}}{n_{kg}^{(1)}}\frac{B((\alpha_0+n_{kg}^{(1)}),(\beta_0+n_{kg}^{(0)}))}{B(\alpha_0,\beta_0)} \\ &= \frac{1}{(n_{kg}+1)}\frac{1}{B((n_{kg}^{(0)}+1),(n_{kg}^{(1)}+1))}\frac{B((\alpha_0+n_{kg}^{(1)}),(\beta_0+n_{kg}^{(0)}))}{B(\alpha_0,\beta_0)} \end{aligned}$$

I.e., the Beta-binomial distribution.

4. Set $\boldsymbol{\theta} = (a, \alpha_0, \beta_0, \sigma_0, \nu_0)$, then from (C.8), (3) and (6), we have the marginal

distribution of $\hat{\beta}_{kg}^{(M)}$ as

$$\hat{\beta}_{kg}^{(M)}|\boldsymbol{\theta} \sim \sum_{n_{kg}^{(1)}} f^*_{n_{kg}^{(1)}}(n_{kg}^{(1)}) \mathcal{G}_{\tau_{kg}, s_{kg}^2}(\sigma_0, \nu_0, a) \tag{C.9}$$

where $\mathcal{G}_{\tau_{kg}, s_{kg}^2}(\sigma_0, \nu_0, a)$ has the form given in (6) and shown below

$$\frac{\exp\left(-\frac{(\hat{\beta}_{kg}^{(M)})^2}{2(a+s_{kg}^2)}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)\sqrt{2\pi(a+s_{kg}^2)}} \left(\frac{\nu_0\sigma_0^2\tau_{kg}}{2(a+s_{kg}^2)}\right)^{\nu_0/2} \Gamma\left(\frac{\nu_0+1}{2}\right) \times$$

$$\sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{(\hat{\beta}_{kg}^{(M)})^2}{2(a+s_{kg}^2)}\right)^i \cdot U\left(\frac{\nu_0+1}{2}, \frac{\nu_0}{2}+1-i, \frac{\nu_0\sigma_0^2\tau_{kg}}{2(a+s_{kg}^2)}\right)$$

and $\tau_{kg} = \dfrac{n_{kg}^{(1)}}{n_{kg}}\ell_{kg}$.

Case 2; $n_{kg}^{(1)} = 0$

Here

$$\begin{aligned}
\int_{\pi_g}\int_{\sigma_g^2} \left(\hat{\beta}_{kg}^{(M)}|\pi_g, \sigma_g^2\right) f_{\sigma_g^2,\pi_g}(\sigma_g^2, \pi_g) &= \int_{\pi_g}\int_{\sigma_g^2} \left[f_{N_{kg}^{(1)}|\pi_g}(0)\mathcal{N}\left(0, a+s_{kg}^2\right)\right] f_{\sigma_g^2,\pi_g}(\sigma_g^2, \pi_g) \\
&= \int_{\pi_g} \left[f_{N_{kg}^{(1)}|\pi_g}(0)\mathcal{N}\left(0, a+s_{kg}^2\right)\right] f_{\pi_g}(\pi_g) \\
&= f^*_{n_{kg}^{(1)}}(0)\mathcal{N}\left(0, a+s_{kg}^2\right)
\end{aligned}$$

Hence the likelihood for snp k with respect to gene g is

$$\mathcal{L}(\theta, \hat{\beta}_{kg}^{(M)}) = f^*_{n_{kg}^{(1)}}(0)\mathcal{N}\left(0, a + s_{kg}^2\right) + \sum_{n_{kg}^{(1)} \geq 1} f^*_{n_{kg}^{(1)}}(n_{kg}^{(1)})\mathcal{G}_{\tau_{kg}, s_{kg}^2}(\sigma_0, \nu_0, a)$$

note that

$$E(\hat{\beta}_{kg}^{(M)}) = 0 \text{ and } \text{Var}(\hat{\beta}_{kg}^{(M)}) = a + s_{kg}^2 + \tau_{kg}\sigma_0^2 \frac{\nu_0}{\nu_0 - 2} \text{ for } \nu_0 > 2$$

5. Ignoring correlation across genes and between snps we use the composite likelihood under a working independence assumption. This is given as

$$\mathbf{CL}(\hat{\beta}^{(M)}|\boldsymbol{\theta}) = \prod_{g=1}^{G}\prod_{k=1}^{K_g} \mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kg}^{(M)}) \tag{C.10}$$

Hence the composite log likelihood is given as

$$\begin{aligned}\mathbf{cl}(\hat{\beta}^{(M)}) &= \sum_{g=1}^{G}\sum_{k=1}^{K_g} \log\left(f^*_{n_{kg}^{(1)}}(0)\mathcal{N}\left(0, a + s_{kg}^2\right) + \sum_{n_{kg}^{(1)} \geq 1} f^*_{N_{kg}^{(1)}}(n_{kg}^{(1)})\mathcal{G}_{\tau_{kg}, s_{kg}^2}(\sigma_0, \nu_0, a)\right) \\ &= \sum_{g=1}^{G}\sum_{k=1}^{K_g} \log\left(\mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kg}^{(M)})\right)\end{aligned}$$

We estimate 5 parameters, I.e., $\boldsymbol{\theta} = (\alpha_0, \beta_0, \sigma_0, \nu_0, a)$.

Optimization

We can see from (C.8) that the likelihood is maximized over a large set of points. To optimize this, we propose the following simplifications using a sieve approach.

We partition the likelihood over the genome by defining neighboring disjoint sets $S_r, r = 1 \ldots, R$ such that

$$\mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kg}^{(M)}) = \mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kr}^{(M)}) \text{ for all } g \in S_r$$

Let $C_r = |S_r|$, that is, C_r is the number of elements in S_r and note that S_r is defined so that $\sum_{r=1}^{R} |S_r| = \sum_{r=1}^{R} C_r = G$. For each set, S_r, we define $\hat{\beta}_{kr}^{(M)}$ such that[2]

$$\frac{\sum_{g \in S_r} I(\hat{\beta}_{kg}^{(M)} \leq \hat{\beta}_{kr}^{(M)})}{C_r} = \frac{\sum_{g \in S_r} I(\hat{\beta}_{kg}^{(M)} > \hat{\beta}_{kr}^{(M)})}{C_r} = .5$$

Hence (C.8) is rewritten as

$$\mathbf{CL}(\hat{\beta}^{(M)}|\boldsymbol{\theta}) = \prod_{g=1}^{G} \prod_{k=1}^{K_g} \mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kg}^{(M)}) = \prod_{k=1}^{K} \prod_{r=1}^{R} \left[\mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kr}^{(M)})\right]^{C_r} \tag{C.11}$$

If $R = G$, I.e., each set contains only one gene, then we obtain (C.8). We can choose R such that $R \ll G$.

[2]Note that this is done on a per SNP basis. I.e., k is fixed.

Estimation

We use the method of differential evolution[106] to obtain the parameter estimates, $\hat{\boldsymbol{\theta}}$, which are global maximizers of the composite log-likelihood with the constraint that $P(h_g^2 \leq 1) = 1$. The reason for this is as follows. Assuming that the gth gene and all the M SNPs in the genotype matrix for the joint model (eqn C.1) have been transformed to have unit variances and mean zero then

$$h_g^2 = \sum_{k=1}^{M} Var(\beta_{kg}^{(J)}) = M\pi_g\sigma_g^2$$

Based on our specification for the distributions of π_g and σ_g^2 in (4) we have that

$$h_g^2|M, \pi_g \sim \text{Inverse Gamma}(a_0, M\pi_g b_0)$$

Hence the density for h_g^2 in our model is

$$f(h_g^2; \boldsymbol{\theta}) = \int_0^1 f(h_g^2|\pi_g; \nu_0, \sigma_0^2) f(\pi_g; \alpha_0, \beta_0) \quad d\pi_g, \quad 0 \leq h_g^2 < \infty \qquad \text{(C.12)}$$

So that based on our model h_g^2 can be greater than 1. To remedy this, we require that $0 \leq h_g^2 \leq 1$ for all genes. Hence, since $h_1^2, \ldots, h_G^2$ are iid from the density in (C.12), an equivalent requirement is $\prod_{g=1}^{G} P(h_g^2 \leq 1) = 1$ or $P(h_g^2 \leq 1) = 1$ for any gene g. Where

$$P(h_g^2 \leq 1) = \int_0^1 f(h_g^2; \boldsymbol{\theta}) \quad dh_g^2$$

Since $P(h_g^2 \leq 1) < 1$ implies that given parameter values $\boldsymbol{\theta}$ there exists some genes with $h_g^2 > 1$ which is not possible.

Estimating $E(h_g^2), Var(h_g^2), E(\pi_g), Var(\pi_g), E(\sigma_g^2),$ **and** $Var(\sigma_g^2)$

Recall that the elements of $\hat{\boldsymbol{\theta}}$ are the paramters for the beta and the inverse gamma distributions. Hence

$$Var(\beta_{kg}^{(J)}) = \pi_g \sigma_g^2$$

is the per SNP heritability for gene g. Hence, the total narrow sense heritability (assuming Y_g is scaled to have unit variance) for gene g is

$$h_g^2 = \sum_{k=1}^{M} Var(\beta_{kg}^{(J)}) = M\pi_g\sigma_g^2$$

With this in mind we can specify the average heritability and its variance across genes and use this as follows.

$$E(h_g^2) = E(M\pi_g\sigma_g^2) = ME(\sigma_g^2 E(\pi_g|\sigma_g^2))$$

$$= M\mu_\pi E(\sigma_g^2)$$

where $\mu_\pi = E(\pi_g)$. Where M is the total number of SNPs, and the variance

$$Var(h_g^2) = Var(M\pi_g\sigma_g^2) = M^2\left[Var(\sigma_g^2 E(\pi_g|\sigma_g^2)) + E((\sigma_g^2)^2 Var(\pi_g|\sigma_g^2))\right]$$

$$= M^2\left[Var(\sigma_g^2)\mu_\pi^2 + E((\sigma_g^2)^2)Var(\pi_g)\right]$$

Where $E(\pi_g), Var(\pi_g), E(\sigma_g^2)$, and $Var(\sigma_g^2)$ are obtained using the densities in (4)

Variance calculation

Using the same approach seen in [105, 107] the variance for $\hat{\boldsymbol{\theta}}$ is given as

$$\text{var}(\hat{\boldsymbol{\theta}}) = I^{-1}(\boldsymbol{\theta})J(\boldsymbol{\theta})I^{-1}(\boldsymbol{\theta})$$

where

$$I(\boldsymbol{\theta}) = E\left[\sum_{g=1}^{G}\sum_{k=1}^{K_g}\frac{U_{kg}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right], \quad J(\boldsymbol{\theta}) = \text{var}\left\{\sum_{g=1}^{G}\sum_{k=1}^{K_g}U_{kg}(\boldsymbol{\theta})\right\}, \quad U_{kg}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\log\left(\mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kg}^{(M)})\right)$$

We can estimate $I(\boldsymbol{\theta})$ empirically even for correlated data. Hence for $I(\boldsymbol{\theta})$ its empirical estimate at the true value is

$$\hat{I}(\boldsymbol{\theta}) = \sum_{g=1}^{G}\sum_{k=1}^{K_g}\frac{U_{kg}}{\partial\boldsymbol{\theta}} = \sum_{g=1}^{G}\sum_{k=1}^{K_g}\frac{l_{kg}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \quad \boxed{l_{kg}(\boldsymbol{\theta}) = \log\left(\mathcal{L}(\boldsymbol{\theta}, \hat{\beta}_{kg}^{(M)})\right)}$$

Doing the same for $J(\boldsymbol{\theta})$ we have its empirical variance estimate at the truth as

$$\hat{J}(\boldsymbol{\theta}) = \text{Var}\left(\sum_{g=1}^{G}\sum_{k=1}^{K_g} U_{kg}(\boldsymbol{\theta})\right)$$

Both estimates would be consistent. However, we don't have $\boldsymbol{\theta}$ so we use the plug in estimate $\hat{\boldsymbol{\theta}}$, so that

$$\hat{I}(\hat{\boldsymbol{\theta}}) = \sum_{g=1}^{G}\sum_{k=1}^{K_g} \frac{l_{kg}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

$$\hat{J}(\hat{\boldsymbol{\theta}}) = \text{Var}\left(\sum_{g=1}^{G}\sum_{k=1}^{K_g} U_{kg}(\hat{\boldsymbol{\theta}})\right) = E\left(\left\{\sum_{g=1}^{G}\sum_{k=1}^{K_g} U_{kg}(\hat{\boldsymbol{\theta}})\right\}\left\{\sum_{g=1}^{G}\sum_{k=1}^{K_g} U_{kg}(\hat{\boldsymbol{\theta}})\right\}^T\right)$$

We can rewrite the last equation as

$$E\left(\left\{\sum_{g=1}^{G}\sum_{k=1}^{K_g} U_{kg}(\hat{\boldsymbol{\theta}})\right\}\left\{\sum_{g=1}^{G}\sum_{k=1}^{K_g} U_{kg}(\hat{\boldsymbol{\theta}})\right\}^T\right) = E\left(\sum_{i,j,m,n=1}^{N} U_{kg_{i,m}}(\hat{\boldsymbol{\theta}})U_{kg_{j,n}}(\hat{\boldsymbol{\theta}})^T\right)$$

Where $N = \sum_{g=1}^{G} K_g$ is the total number of SNP $\times$ Gene *trans* pairs. Hence, we see that estimating $\hat{J}(\hat{\boldsymbol{\theta}})$ is computationally intensive. To address this we utilize the moving block bootstrap approach. We observe that the correlation between the score statistics for any two Genes and SNPs is non-zero if there is both correlation across Genes for a given SNP and across SNPs for a given Gene. Hence, we sum the score statistic across Genes for a given SNP, and apply the moving block approach across SNPs only. We construct overlapping

(moving) blocks of size L. For each block, we sum across SNPs that fall within it, then sum the results across all blocks. The bootstrap approach is due to the fact that we change the starting point of the moving blocks for each of the B bootstrap samples. The block window size is selected by choosing the maximum number of SNPs tagged by any SNP genome-wide. We then estimate $\hat{J}(\hat{\boldsymbol{\theta}})$ as the observed variance in the bootstrap samples.

Based on transforms

Recall that the elements of $\hat{\boldsymbol{\theta}}$ are the parameters for the Beta and the Inverse Gamma distributions. Using these parameters we can estimate the uncertainty in the estimates of $E(h_g^2), Var(h_g^2), E(\pi_g), Var(\pi_g), E(\sigma_g^2)$, and $Var(\sigma_g^2)$. This is done through two applications of the Delta method. Define

$$g(\boldsymbol{\theta}) = \begin{pmatrix} E(\pi_g) & Var(\pi_g) & E(\sigma_g^2) & Var(\sigma_g^2) \end{pmatrix}$$

Hence $g(\cdot)$ is a function of α_0, β_0, ν_0, and σ_0^2. Hence the covariance matrix for the joint distribution of $E(\pi_g), Var(\pi_g), E(\sigma_g^2)$, and $Var(\sigma_g^2)$ is

$$\Sigma = g'(\boldsymbol{\theta}) var(\boldsymbol{\theta}) (g'(\boldsymbol{\theta}))^T$$

where

$$g'(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial E(\pi_g)}{\partial \alpha_0} & \frac{\partial E(\pi_g)}{\partial \beta_0} & \frac{\partial E(\pi_g)}{\partial \nu_0} & \frac{\partial E(\pi_g)}{\partial \sigma_0^2} \\ \frac{\partial Var(\pi_g)}{\partial \alpha_0} & \frac{\partial Var(\pi_g)}{\partial \beta_0} & \frac{\partial Var(\pi_g)}{\partial \nu_0} & \frac{\partial Var(\pi_g)}{\partial \sigma_0^2} \\ \frac{\partial E(\sigma_g^2)}{\partial \alpha_0} & \frac{\partial E(\sigma_g^2)}{\partial \beta_0} & \frac{\partial E(\sigma_g^2)}{\partial \nu_0} & \frac{\partial E(\sigma_g^2)}{\partial \sigma_0^2} \\ \frac{\partial Var(\sigma_g^2)}{\partial \alpha_0} & \frac{\partial Var(\sigma_g^2)}{\partial \beta_0} & \frac{\partial Var(\sigma_g^2)}{\partial \nu_0} & \frac{\partial Var(\sigma_g^2)}{\partial \sigma_0^2} \end{pmatrix}$$

With this we can then obtain the covariance matrix for $E(h_g^2)$, and $Var(h_g^2)$. Let

$$\boldsymbol{\xi} = g(\boldsymbol{\theta}) = \begin{pmatrix} \mu_{\pi_g} & \tau_{\pi_g} & \mu_{\sigma_g^2} & \tau_{\sigma_g^2} \end{pmatrix}$$

where $\mu_{\pi_g}, \tau_{\pi_g}, \mu_{\sigma_g^2}$, and $\tau_{\sigma_g^2}$ represent $E(\pi_g), Var(\pi_g), E(\sigma_g^2)$, and $Var(\sigma_g^2)$ respectively using $\boldsymbol{\theta}$. Furthermore, let

$$h(\boldsymbol{\xi}) = \begin{pmatrix} E(h_g^2) & Var(h_g^2) \end{pmatrix}$$

Hence the covariance matrix for $E(h_g^2)$, and $Var(h_g^2)$ is

$$\Xi = h'(\boldsymbol{\xi}) \Sigma (h'(\boldsymbol{\xi}))^T$$

where

$$h'(\boldsymbol{\xi}) = \begin{pmatrix} \frac{\partial E(h_g^2)}{\partial \mu_{\pi_g}} & \frac{\partial E(h_g^2)}{\partial \tau_{\pi_g}} & \frac{\partial E(h_g^2)}{\partial \mu_{\sigma_g^2}} & \frac{\partial E(h_g^2)}{\partial \tau_{\sigma_g^2}} \\ \frac{\partial Var(h_g^2)}{\partial \mu_{\pi_g}} & \frac{\partial Var(h_g^2)}{\partial \tau_{\pi_g}} & \frac{\partial Var(h_g^2)}{\partial \mu_{\sigma_g^2}} & \frac{\partial Var(h_g^2)}{\partial \tau_{\sigma_g^2}} \end{pmatrix}$$

C.3 Simulation approach

We generate the summary level results using the following model and simulation scheme. The model used is

$$\hat{\beta}_{kg}^{(M)} = \beta_{kg}^{(M)} + \xi_k + e_{kg}$$

where $\xi_k, k = 1, \ldots, K$ are i.i.d $\mathcal{N}(0, a)$, and the error term $e_g = (e_1, \ldots, e_K)$ follows a multivariate normal distribution with mean zero and covariance matrix $\mathbf{R}$/n. $\mathbf{R}$ is a matrix of LD coefficients for the eQTLs of a given gene, and n being the sample size of the eQTL study with respect to the tissue being used.

For each simulation, we first generate π_g and σ_g^2 for gene g according to (4) using pre-specified values of α_0, β_0, ν_0 and σ_0^2. We then generate $\beta_{kg}^{(J)}$ according to the model in (C.2); from this obtain $\beta_{kg}^{(M)}$ using the result in (C.3). ρ_{kp}, the pairwise Pearson correlation coefficient between markers k and p, is estimated using the corresponding sample correlation coefficient in the reference dataset.

To generate e_g note that $\mathbf{R}$ will be large and that we also need to account for the relationship across genes, so we do the following. We note that the distribution of e_g is the same as the joint distribution of the eQTL summary level statistic under the null of no association between any of the SNPs and the gene. We also note that there is some correlation of the effect sizes expected of given SNP across genes, so we do the following.

1. Using the n_{ref} subjects in our 1000 GENOME reference dataset. We generate, independently, standardized pseudo genes $\mathbf{Y}_i = (Y_1, \ldots, Y_G), i = 1, \ldots, n_{ref}$

from a multivariate normal distribution. I.e

$$\mathbf{Y}_i \sim \text{MVN}(0, V)$$

where V can either be estimated from a reference dataset to account for correlation between the G genes or assumed to be an identity matrix to represent independence across the G genes.

2. Using the genotype data in the reference panel, we calculate the standardized effect sizes for a given gene $\mathbf{u}_g = (u_1, \ldots, u_k)$ for the K SNPs.

3. Set $e_{kg} = u_{kg}\sqrt{n_{ref}/n}$ to account for the difference in sample size between the reference dataset and the eQTL study. Since $\mathbf{u}_g \sim \mathcal{N}(0, \mathbf{R}/n_{ref})$, then $\mathbf{e}_g \sim \mathcal{N}(0, \mathbf{R}/n)$

Addendum

We use the following transformation

$$\sigma_g^2 = \frac{h_g^2}{M\pi_g}$$

this implies that the distribution of $h_g^2|\pi_g$ as

$$h_g^2|\pi_g \sim \text{Inv-Gamma}(a, M\pi_g b)$$

Hence, we generate π_g first then using this as well the number of SNPs we then generate h_g^2 from the distribution above. Using the transformation

$$\sigma_g^2 = \frac{h_g^2}{M\pi_g}$$

we obtain σ_g^2, with each following the specification in (4).

C.4 Future projection

OWith the parameters estimated above, we can provide estimates of the future yield for other studies using the observed effect sizes. Let ND_α be the number of Gene-SNP associations obtained at the type 1 error rate of α[3]. Furthermore, assuming that both the expression levels per Gene and allele count per SNP have been transformed to have unit variances and mean zero, then using the per Gene joint effect sizes, $\hat{\beta}_{sg}^{(J)}$ for SNP s and Gene g, we have

$$\begin{aligned} E(ND_\alpha) &= E(E(ND_\alpha|Gene)) \\ &= \sum_{g=1}^{G}\sum_{s=1}^{M} P\left(\sqrt{n} \mid \hat{\beta}_{sg}^{(J)} \mid > z_{\alpha/2} \left|\beta_{sg}^{(J)}\right.\right) \\ &\approx G * M \int_{\sigma_g^2\times\pi_g}\int_{\beta_{sg}^{(J)}} \text{pow}_\alpha(\beta_{sg}^{(J)}) p(\beta_{sg}^{(J)}|\pi_g,\sigma_g^2) p(\pi_g,\sigma_g^2;\hat{\boldsymbol{\theta}}) d\beta_{sg}^{(J)} d\pi_g d\sigma_g^2 \end{aligned}$$

[3]To account for multiple testing using the Benjamini-Hochberg approach, this is $\alpha/(\#SNPs * \#Genes)$

Where $\text{pow}_\alpha(\beta) = \Phi(-z_{\alpha/2} - \beta\sqrt{n}) + 1 - \Phi(z_{\alpha/2} - \beta\sqrt{n})$, $\Phi(\cdot)$ is cumulative distribution function for the standard normal distribution, $z_\alpha = \Phi(1-\alpha)$ is the αth quantile of the standard normal distribution, and $p(\beta_{sg}^{(J)}|\pi_g, \sigma_g^2)p(\pi_g, \sigma_g^2; \hat{\boldsymbol{\theta}})$ is the inferred effect size distribution for a given gene. Using the normal-mixture model with the inverse gamma and beta priors per gene, we have the following result after marginalizing over the gene specific priors

$$E(ND_\alpha) \approx G * M * E(\pi_g) \int_\beta \text{pow}_\alpha(\beta) t(\beta; \sigma_0, \nu_0) d\beta.$$

where $t(\cdot)$ is the generalized student distribution with location parameter zero, scale parameter σ_0, and ν_0 degress of freedom.

In a similar vein we can obtain the expected value of the proportion of genetic variance explained by susceptibility SNPs reaching genome-wide significance, after accounting for multiple testing and averaged across genes, as

$$\begin{aligned} E(GV_\alpha) &= E(E(GV_\alpha|Gene)) \\ &\approx \int_{\sigma_g^2} \sigma_g^2 \int_\beta \beta^2 \text{pow}_\alpha(\beta) \mathcal{N}(0, \sigma_g^2) p(\sigma_g^2) d\beta d\sigma_g^2 \end{aligned}$$

This can be simplified as

$$E(GV_\alpha) \approx \int_\beta \beta^2 \text{pow}_\alpha(\beta) f(\beta; \sigma_0, \nu_0) d\beta$$

where

$$f(\beta;\sigma_0,\nu_0) = \frac{\Gamma\left(\frac{\nu_0+3}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)\sqrt{2\pi}}\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\nu_0/2}\left(\frac{2}{x^2+\nu_0\sigma_0^2}\right)^{(\nu_0+3)/2}$$

C.5 Derivations

1. Show

$$\beta_k^{(M)} = \sum_{p=1}^{P}\beta_p^{(J)}\rho_{kp}$$

where ρ_{kp} is the correlation coefficient between snps k and p.

Under the polygenic model for the gth gene expression level we have

$$Y_g = X\beta^{(J)} + \epsilon$$

where Y_g is a n $\times$ 1 vector of gene expression levels and X is the n $\times$ P matrix of genotypes for P SNPs and n subjects. Furthermore, assume that both Y_g and X have both been transformed to have a sample variance of 1 and a sample mean of 0. Then the OLS estimate, $\hat{\beta}_k^{(M)}$, for the kth snp is derived as

$$\hat{\beta}_k^{(M)} = (X_k^T X_k)^{-1}X_k^T Y_g = \frac{1}{n}X_k^T Y_g$$

This can be further simplified based on the polygenic model as

$$\hat{\beta}_k^{(M)} = \frac{1}{n}X_k^T(X\hat{\beta}^{(J)} + \epsilon) = \frac{1}{n}\sum_{p=1}^{P}X_k^T X_p\hat{\beta}_p^{(J)} + 0$$

since the columns of X are transformed to have sample variances of 1 and sample means of 0, then $X_k^T X_p = n\rho_{kp}$. Furthermore $X_k^T \epsilon = 0$ since X_k is in the column space of X.

2. Let $Y|\mu \sim \mathcal{N}(\mu, \sigma^2)$ and μ be t-distributed with location parameter zero and scale parameter k with ν degrees of freedom, then the distribution of Y marginalized over μ is given as

$$f_Y(y) = \frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi\sigma^2}} \left(\frac{\nu k^2}{2\sigma^2}\right)^{\nu/2} \exp\left(\frac{\nu k^2}{2\sigma^2}\right) \int_1^\infty q^{-\frac{1}{2}} (q-1)^{(\nu-1)/2} \exp\left(\frac{1}{2\sigma^2}\left[\frac{y^2}{q} - q\nu k^2\right]\right)\, dq$$

Derivation

Recall that

$$\Gamma(s) = \int_0^\infty t^{s-1} \exp(-t)\, dt$$

Hence

$$\Gamma\left(\frac{\nu+1}{2}\right)\left(1 + \frac{1}{\nu}\left[\frac{\mu}{k}\right]^2\right)^{-\left(\frac{\nu+1}{2}\right)} = \int_0^\infty w^{\left(\frac{\nu+1}{2}\right)-1} \exp\left(-w\left[1 + \frac{1}{\nu}\left(\frac{\mu}{k}\right)^2\right]\right)\, dw \qquad (A)$$

Therefore

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|\mu}(y) \cdot f_\mu(\mu)\ d\mu$$

$$= \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi k^2 \nu}} \left(1 + \frac{1}{\nu}\left(\frac{\mu}{k}\right)^2\right)^{-\left[\frac{\nu+1}{2}\right]} d\mu$$

$$= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu k^2}} \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{1}{2}\left[\frac{y-\mu}{\sigma}\right]^2\right)}{\sqrt{2\pi\sigma^2}} \times$$

$$\int_0^{\infty} w^{\left(\frac{\nu+1}{2}\right)-1} \exp\left(-w\left[1 + \frac{1}{\nu}\left(\frac{\mu}{k}\right)^2\right]\right)\ dw d\mu \qquad \boxed{\text{From (A)}}$$

This simplifies to

$$\frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu k^2}} \int_0^{\infty} \left(1 + \frac{2w\sigma^2}{\nu k^2}\right)^{-1/2} w^{\left(\frac{\nu+1}{2}\right)-1} \exp(-w) \exp\left(\frac{y^2}{2\sigma^2}\left[1 + \frac{2w\sigma^2}{\nu k^2}\right]^{-1}\right) dw \quad (B)$$

setting $q = 1 + \frac{2w\sigma^2}{\nu k^2}$, (B) becomes

$$f_Y(y) = \frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi\sigma^2}} \left(\frac{\nu k^2}{2\sigma^2}\right)^{\nu/2} \exp\left(\frac{\nu k}{2\sigma^2}\right) \int_1^{\infty} q^{-\frac{1}{2}} (q-1)^{(\nu-1)/2} \exp\left(\frac{1}{2\sigma^2}\left[\frac{y^2}{q} - q\nu k^2\right]\right) dq$$

For a specific value of ν, we obtain the form given in 6. More generally, if we

use the transformation $t = \dfrac{2w\sigma^2}{\nu k^2}$ instead, we have (B) simplifying as

$$f_Y(y) = \frac{\exp\left(-\dfrac{y^2}{2\sigma^2}\right)}{\Gamma\left(\dfrac{\nu}{2}\right)\sqrt{\pi\nu k^2}}\left(\frac{\nu k^2}{2\sigma^2}\right)\left(\frac{\nu+1}{2}\right)\int_0^\infty \frac{t^{[(\nu+1)/2]-1}}{\sqrt{1+t}}\exp\left(\frac{1}{2\sigma^2}\left[\frac{y^2}{1+t} - t\nu k^2\right]\right)\,dt \tag{C.13}$$

Simplifying the integral we have

$$\int_0^\infty \frac{t^{[(\nu+1)/2]-1}}{\sqrt{1+t}}\exp\left(\frac{1}{2\sigma^2}\left[\frac{y^2}{1+t} - t\nu k^2\right]\right)\,dt = \int_0^\infty \frac{t^{[(\nu+1)/2]-1}}{\sqrt{1+t}}\exp\left(-\frac{t\nu k^2}{2\sigma^2}\right)\times \sum_{i=0}^{\infty}\left(\frac{y^2}{2\sigma^2(1+t)}\right)^i/i!\;dt$$

Recall that Tricomi's (confluent hypergeometric) function, U(a,b,z) is given as

$$U(a,b,z) = \frac{1}{\Gamma(a)}\int_0^\infty \exp(-zt)t^{a-1}(1+t)^{b-a-1}\,dt,\ \text{(Re a} > 0\ )$$

Setting $a = \dfrac{\nu+1}{2}$, $b = \dfrac{\nu}{2} + 1 - i$, and $z = \dfrac{\nu k^2}{2\sigma^2}$ we have

$$\int_0^\infty \frac{t^{[(\nu+1)/2]-1}}{\sqrt{1+t}}\exp\left(\frac{1}{2\sigma^2}\left[\frac{y^2}{1+t} - t\nu k^2\right]\right)\,dt = \Gamma\left(\frac{\nu+1}{2}\right)\sum_{i=0}^{\infty}\frac{1}{i!}\left(\frac{y^2}{2\sigma^2}\right)^i \times U\left(\frac{\nu+1}{2}, \frac{\nu}{2}+1-i, \frac{\nu k^2}{2\sigma^2}\right)$$

Hence,

$$f_Y(y) = \frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi\sigma^2}} \left(\frac{\nu k^2}{2\sigma^2}\right)^{\nu/2} \Gamma\left(\frac{\nu+1}{2}\right) \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{y^2}{2\sigma^2}\right)^i \cdot U\left(\frac{\nu+1}{2}, \frac{\nu}{2}+1-i, \frac{\nu k^2}{2\sigma^2}\right) \tag{C.14}$$

Furthermore, from (B) we have

$$\int_{-\infty}^{\infty} f_Y(y)dy = \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{y^2}{2\sigma^2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu k^2}} \int_0^{\infty} \left(1+\frac{2w\sigma^2}{\nu k^2}\right)^{-1/2} w^{\left(\frac{\nu+1}{2}\right)-1} \times$$

$$\exp(-w)\exp\left(\frac{y^2}{2\sigma^2}\left[1+\frac{2w\sigma^2}{\nu k^2}\right]^{-1}\right)\, dwdy$$

$$= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \int_0^{\infty} w^{\frac{\nu}{2}-1} \exp(-w)\, dw$$

$$= 1 \qquad \boxed{\Gamma(s) = \int_0^{\infty} t^{s-1}\exp(-t)\, dt}$$

C.6 Supplementary Figures

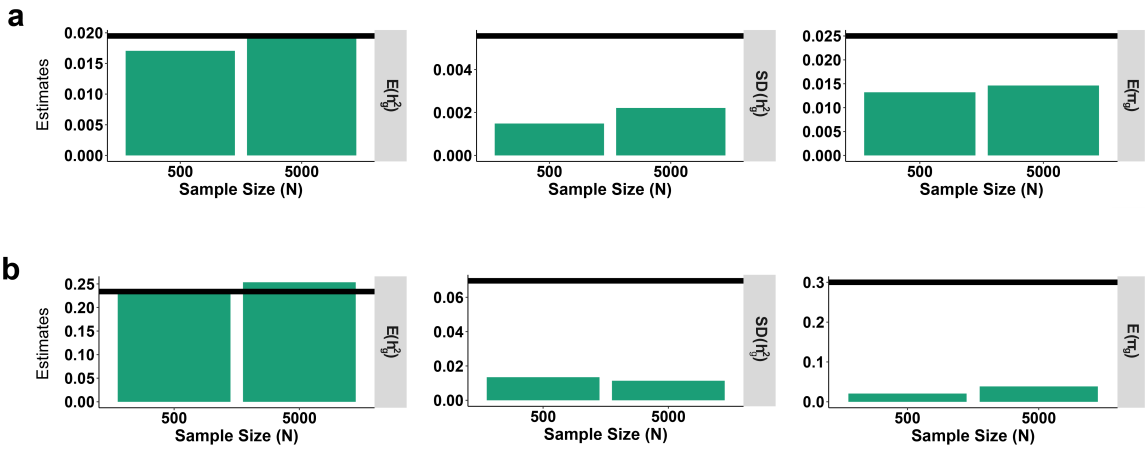


Fig. C.1. Comparison of estimates obtained averaged across 50 datasets at a large per SNP heritability (5e-5). We show results based on a true average polygenicity of 2.5% (**a**) and a true average polygenicity of 30% (**b**). The horizontal black lines correspond to the truth. Note that the y-axis in each subplot are in different scales.

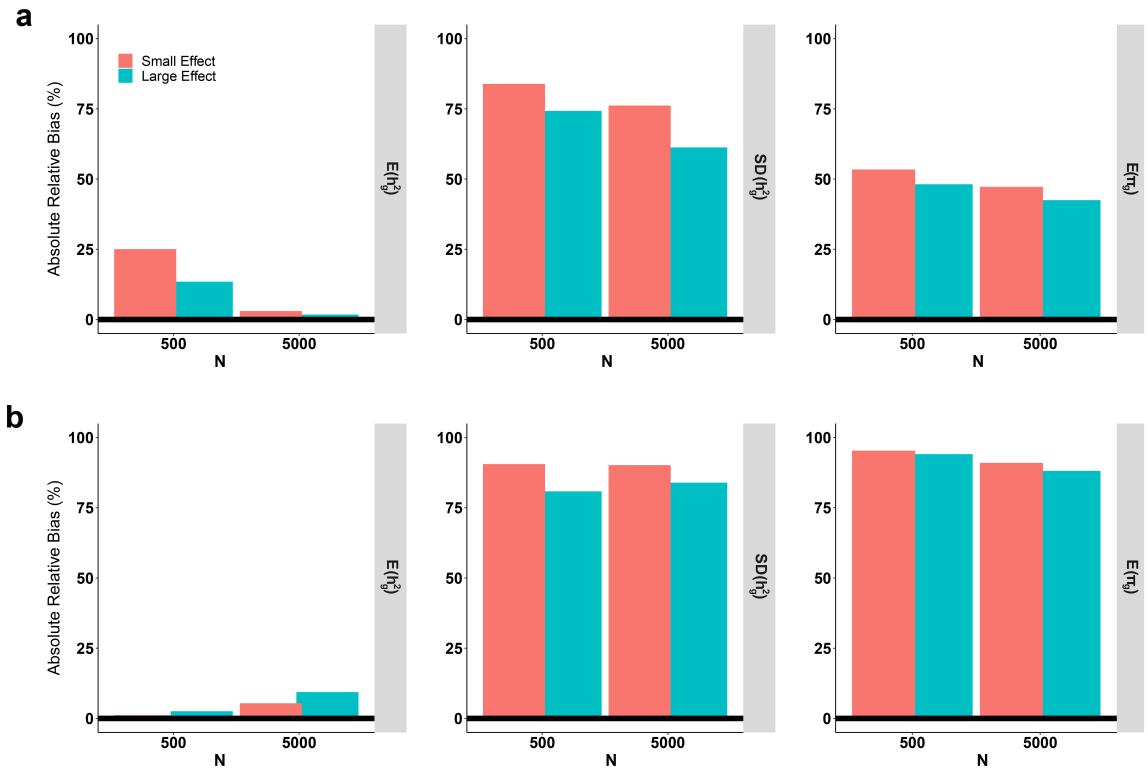


Fig. C.2. Estimated bias obtained averaged across 50 datasets as the per SNP heritability increases (Large effect = 5e-5 vs Small effect = 4e-7). We show results based on a true average polygenicity of 2.5% (**a**) and a true average polygenicity of 30% (**b**). The sample size, N, when the per SNP heritability is small (4e-7) is inflated by a factor of 64. The horizontal black lines correspond to the bias at the truth.

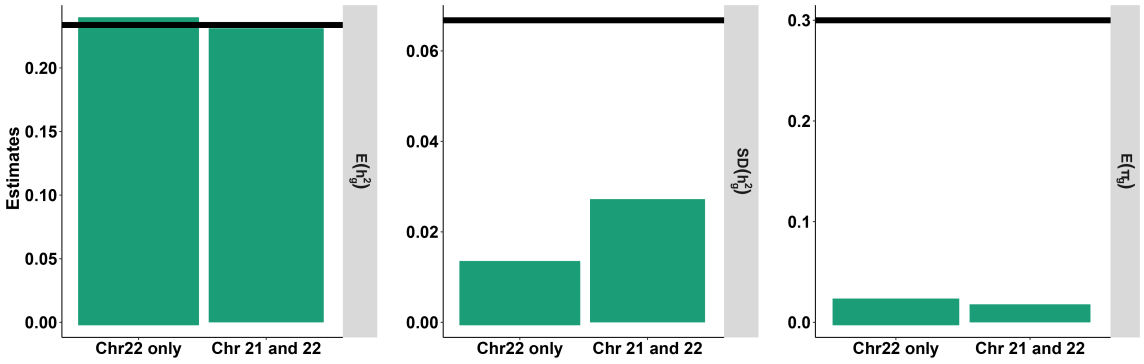


Fig. C.3. Effect of increasing SNP size on estimation at a sample size of 500 when $\mathbf{E}(h_g^2)$, $\mathbf{E}(\pi_g)$, and $\mathbf{SD}(h_g^2)$ are fixed. We show results for $\mathrm{E}(h_g^2)$, $\mathrm{SD}(h_g^2)$, and $\mathrm{E}(\pi_g)$ respectively. Horizontal black lines correspond to the truth. Note that the y-axis in each subplot are in different scales.

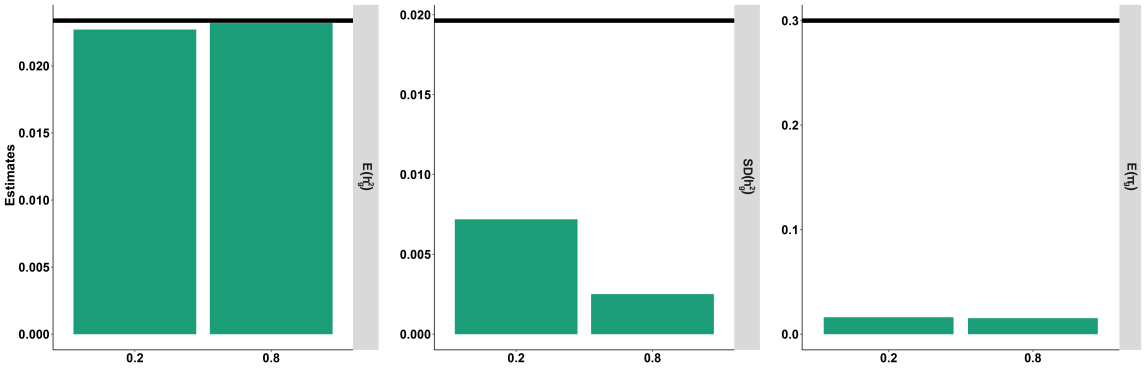


Fig. C.4. Comparison of estimates obtained averaged across 50 datasets at a sample size of 5,000 when $\mathbf{E}(h_g^2)$, $\mathbf{E}(\pi_g)$, and $\mathbf{SD}(h_g^2)$ are fixed. Horizontal black lines correspond to the truth. The plots from left to right are for $\mathrm{E}(h_g^2)$,$\mathrm{SD}(h_g^2)$, and $\mathrm{E}(\pi_g)$ respectively. Note that the y-axis in each subplot are in different scales.

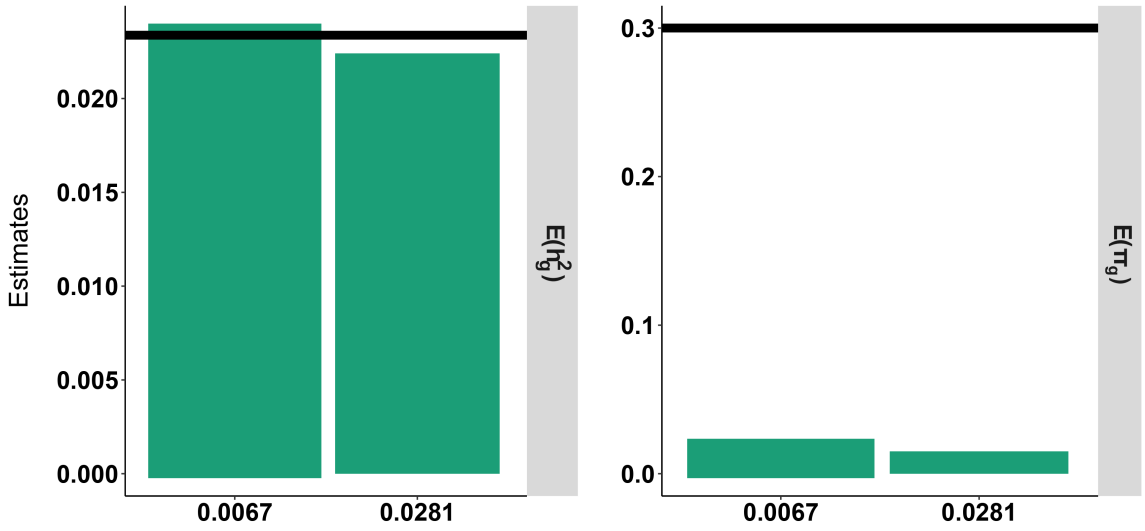


Fig. C.5. Comparison of estimates obtained averaged across 50 datasets at a sample size of 5,000 as SD(h_g^2) increases when E(h_g^2), and E(π_g) are fixed. Horizontal black lines correspond to the truth. Values on the y-axis represent estimates obtained from model fit while values on the x-axis represent SD(h_g^2). The plots from left to right are for E(h_g^2),and E(π_g) respectively. Note that the y-axis in each subplot are in different scales.